# MVQA: Mamba with Unified Sampling for Efficient Video Quality Assessment

Supplementary Material

## 1. Visualization of USDS

In Fig. 1, we show the details of USDS sampling in detail. For each sampled region, we sample three times at the original resolution and splice the sampled three chunks according to the relative positions of the samples, and then scale the sampled region to the same size as the sampled chunks and splice it with the three chunks sampled at the original resolution. Note that the scaled block is always placed in the lower right corner. The sampled map obtained by sampling and splicing not only contains distortion information, but also retains sufficient semantic information, since the lower right block is a scaling of the entire sampled block.

## 2. Experiments on Proportion of Mask

During the sampling masking process, we can set the masking ratio to determine the amount of semantic content to be retained, as shown in Fig. 2. The experimental results with different mask ratios are shown in Tab. 1. When masking at ratios of $1/16$ and $15/16$, the size of the sampled blocks is $8 \times 8$. However, since the input block size of the model is $16 \times 16$, each input block contains four smaller blocks. However, when each $16 \times 16$ input chunk contains four $8 \times 8$ sampling chunks,this heterogeneity destroys the overall consistency of the input chunks, making Patch Embedding unable to effectively integrate the four $8 \times 8$ sampling chunks into a unified feature representation. As a result, the model is inefficient in extracting the local features of each small block, leading to a significant decrease in the ability to characterize distorted and semantic information. Additionally, when the ratio is set to $1/2$ and $3/4$, the performance decreases, which suggests that while semantic information plays an important role in VQA, quality information has a greater impact on quality assessment. As verified in CLiF-VQA [2] using only semantic information for VQA is very ineffective, this is because the distorted information plays a crucial role in the prediction of the model, while the semantic information complements the distorted information. Therefore, when the proportion of semantic information is too large, it leads to the deterioration of the model's effectiveness, so the proportion of semantic information needs to be controlled.

Table 1. Experimental results on proportion of mask.

| Datasets | **LSVQ$_{test}$** | | **KoNViD-1k** | | **LIVE-VQC** | |
|---|---|---|---|---|---|---|
| Proportion | SROCC | PLCC | SROCC | PLCC | SROCC | PLCC |
| $1/16$ | 0.834 | 0.830 | 0.826 | 0.824 | 0.760 | 0.795 |
| $15/16$ | 0.820 | 0.815 | 0.811 | 0.815 | 0.749 | 0.780 |
| $2/4$ | 0.865 | 0.866 | 0.857 | 0.861 | 0.795 | 0.838 |
| $3/4$ | 0.856 | 0.859 | 0.830 | 0.836 | 0.772 | 0.820 |
| $1/4$ | 0.882 | 0.883 | 0.870 | 0.868 | 0.828 | 0.848 |

## 3. More Experiments on Semantic Analysis

In order to fully verify that USDS can effectively retain the semantic information of the video, we conduct a large number of experiments to illustrate this, as shown in Fig. 3. Consistent with the experimental scheme in the paper, we use the CLIP [3] model, which has strong visual language capabilities, for semantic perception. Specifically, we design a textual description for each experimental video that is most relevant to the semantic information of the video, and then this textual description is input into CLIP for semantic perception along with an empty description. Numerous experimental results demonstrate that the USDS designed in this paper can retain more semantic information of the videos compared to Fragments. Specifically, in all the results in Fig. 3, the average semantic score of Fragments is 0.353, while the average score of USDS is 0.845.

## 4. Comparison of USDS and Simple Connection

Compared to the semantic fusion approach of USDS, the simple connection of original video frames with complete semantic information directly to the sampling results is the most intuitive semantic fusion approach. Here, we compare the connected fusion approach with USDS. Specifically, I scale the original video frames to a size of $224 \times 224$ as a semantic map and then concatenate them with the sampled distortion maps from Fragments. We keep the total number of frames in the input of the model to 32 frames. We compare the effect of connecting 1, 4 and 16 frames of the semantic map respectively. As shown in Tab. 2. From the experimental results, it can be concluded that better results can be obtained by USDS compared to simple connection.

Table 2. We perform simple concatenation with the semantic information of 1, 4 and 16 frames, respectively.

| Datasets | **LSVQ$_{test}$** | | **KoNViD-1k** | | **LIVE-VQC** | |
|---|---|---|---|---|---|---|
| Sampling | SROCC | PLCC | SROCC | PLCC | SROCC | PLCC |
| *Concatenation (1)* | 0.850 | 0.852 | 0.838 | 0.840 | 0.766 | 0.802 |
| *Concatenation (4)* | 0.862 | 0.862 | 0.849 | 0.852 | 0.783 | 0.808 |
| *Concatenation (16)* | 0.848 | 0.846 | 0.834 | 0.834 | 0.759 | 0.773 |
| **USDS** | 0.882 | 0.883 | 0.870 | 0.868 | 0.828 | 0.848 |

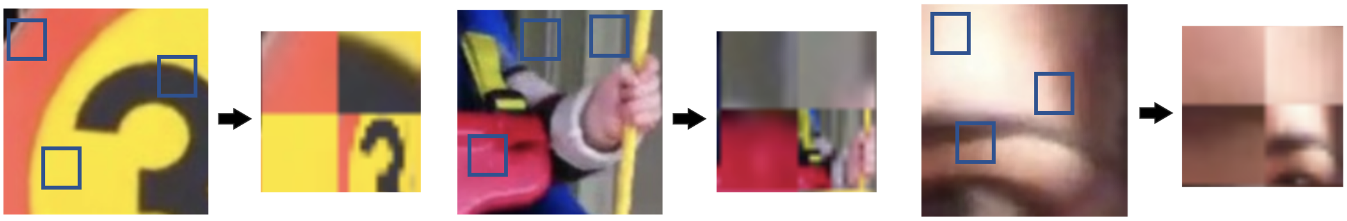

Figure 1. Visualization of USDS.

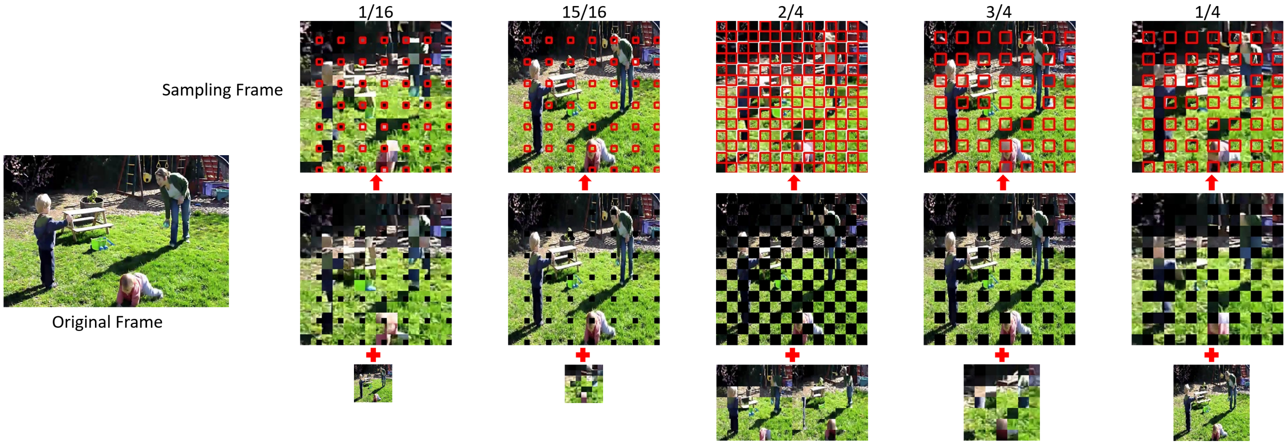

Figure 2. A schematic of the different scale masks. We perform mask fusion between distortion sampling and semantic sampling at different proportions, embedding the smaller sampling results into the larger ones. The red boxes represent the embedding locations.

## 5. Additional Description of Single-branch and Multiple-branches

In the experimental section of the paper, we categorize the deep learning-based VQA methods into single-branch and multi-branch methods. This type of method usually extracts only one type of features, such as spatial details or statistical properties, from the input video data. Single-branch methods rely on a single path and aim to capture core quality-related information in a compact and computationally efficient framework. The computational complexity of single-branch methods is typically low, giving them an advantage in resource-constrained scenarios. However, since video quality perception involves multiple visual and temporal cues, it may be difficult for single-branch methods to fully capture its complexity. In contrast, multi-branch approaches utilize a more complex architecture that contains two or more independent processing channels. Each branch is typically designed to extract specific types of feature information, enabling a richer and more comprehensive analysis of video quality. For example, one branch may focus on extracting spatial features such as texture or edge information, while another branch targets temporal features such as motion coherence or inter-frame consistency. By integrating these heterogeneous feature representations through a fusion mechanism, multibranch approaches are better able to model the complex interactions between multiple factors that affect the perception of video quality. However, this enhanced flexibility and granularity not only increases the complexity of model design, but also significantly increases the computational complexity, making multibranch methods often far more demanding in terms of computational resources than single-branch methods. The difference between these two paradigms highlights a fundamental trade-off in VQA research: single-branch methods achieve efficiency and simplicity with low computational complexity, whereas multi-branch methods trade higher computational complexity for robustness and adaptability to diverse characteristics of video content and distortion. In this study, on the one hand, thanks to the unique characteristics of USDS sampling, our method is able to extract both distortion features and semantic features under a single branch, which retains the low complexity advantage of single-branch methods; on the other hand, by effectively integrating distortion and semantic information, our mamba-based architecture significantly reduces the computational complexity while realizing significant performance improvements compared to previous methods. On the other hand, by effectively integrating the distortion and semantic information, our mamba-based architecture significantly reduces the computational complexity while realizing sig-

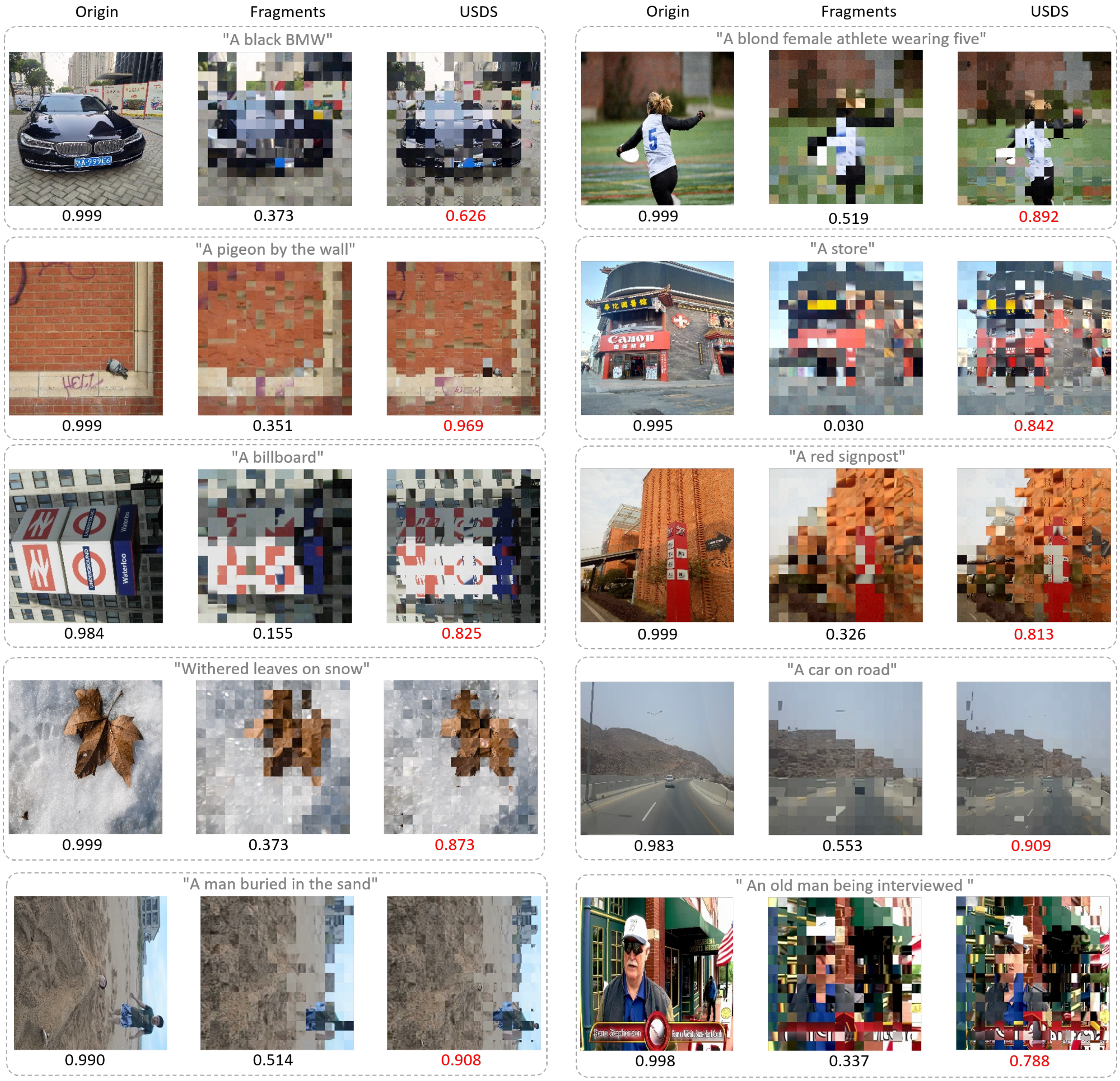

Figure 3. Results of semantic analysis experiments.

nificant performance improvement over previous methods, providing an efficient and powerful solution for video quality evaluation.

## References

[1] Junjie Ke, Tianhao Zhang, Yilin Wang, Peyman Milanfar, and Feng Yang. Mret: Multi-resolution transformer for video quality assessment. *Frontiers in Signal Processing*, 3: 1137006, 2023.

[2] Yachun Mi, Yan Shu, Yu Li, Chen Hui, Puchao Zhou, and Shaohui Liu. CLiF-VQA: Enhancing video quality assessment by incorporating high-level semantic information related to human feelings. In *ACM MM*, page 9989–9998, 2024. 1

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 1

[4] 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*, pages 538–554. Springer, 2022.