MDS-VQA: Model-Informed Data Selection for Video Quality Assessment

Supplementary Material

Content

This supplementary material provides:

- Additional implementation details of the experimental pipeline in Sec. 6;
- Additional qualitative gMAD comparisons against competing methods in Sec. 7;
- Qualitative visualizations of representative failure samples identified by the failure predictor in Sec. 8.

6. Additional Implementation Details

This section supplements Sec. 4.1 by describing the detailed training. We first train VisualQuality-R1 [44] on the YouTube-UGC training set to obtain a base VQA model. For rapid validation under limited compute, we adopt the default LoRA configuration of VLM-R1 [26], using group size $K = 6$, LoRA rank $r = 64$, and a per-GPU batch size of 24, yielding an effective batch size of 48 with 2 gradient accumulation steps. While LoRA can affect the absolute performance of the base model, it does not compromise the comparative validity of our study, because our evaluation focuses on relative improvements across different selection strategies. During training, each video pair is fed to VisualQuality-R1 together with a structured text prompt (see Table 10), producing scalar outputs that are used to compute rewards and optimize the VQA model.

Table 10. Structured text prompt used for training $f(\cdot)$.

You are doing a video quality assessment task. Here is the question: What is your overall rating on the quality of this video? The rating should be a float between 1 and 5, rounded to two decimal places, with 1 representing very poor quality and 5 representing excellent quality. First output the thinking process in `<think> </think>` tags and then output the final answer with only one score in `<answer> </answer>` tags.

7. Additional gMAD Pairs

Figs. 4-6 provide additional representative gMAD pairs to further compare fine-tuned models induced by MDS-VQA against those by ALCS [46], FreeSel [45], and NoiseStability [13]. Across these comparisons, the MDS-VQA-induced model more consistently exposes distinct failure modes of competing methods (when acting as the attacker), and remains more robust under attacks (when acting as the defender), yielding predictions that better agree with human perception of video quality.

8. Visualizations of Failure Samples

Fig. 7 visualizes representative challenging videos selected from YouTube-SFV SDR [37] by the proposed MDS-VQA. Without an explicit diversity constraint, the selection tends to concentrate on visually similar hard samples that share a common failure cause. In Figs. 7(a)-(f), this manifests as many black-toned scenes where the base VQA model predicts less reliably. After incorporating the diversity term, the selected set becomes substantially broader in both semantic content and distortion characteristics (see Figs. 7(g)-(l)), indicating the combined "hard-and-diverse" criterion better covers complementary failure modes under the same labeling budget.

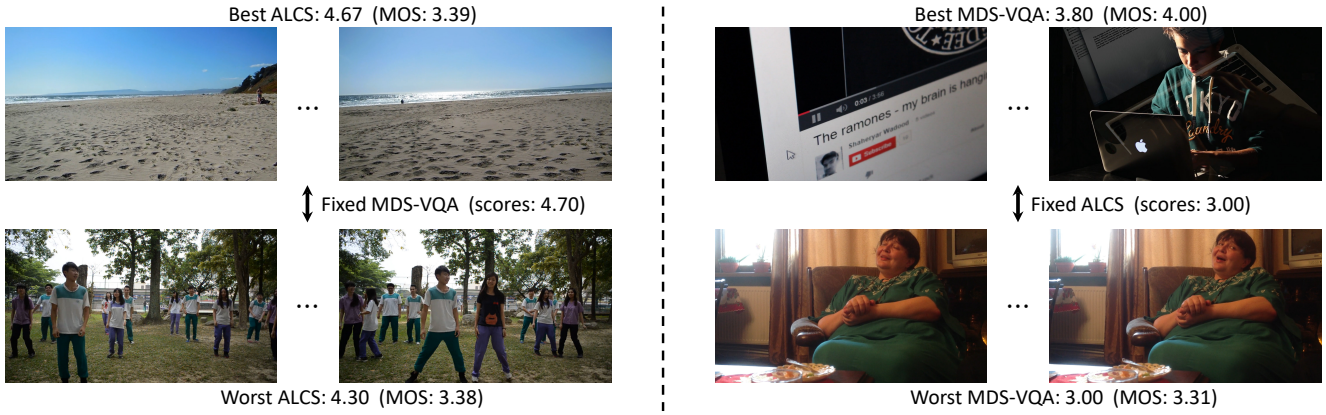

Figure 4. Representative gMAD pairs between VQA models induced by MDS-VQA and ALCS [46].

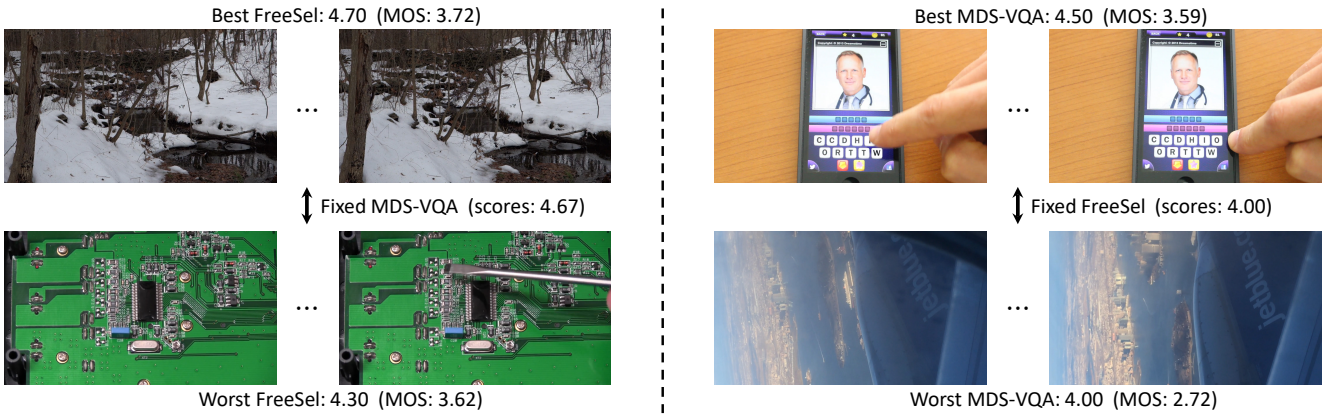

Figure 5. Representative gMAD pairs between VQA models induced by MDS-VQA and FreeSel [45].

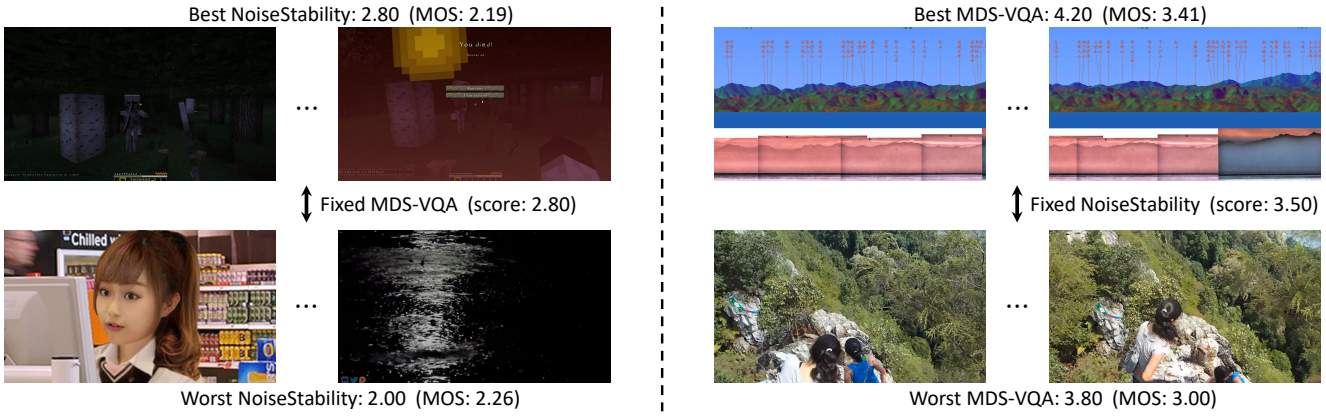

Figure 6. Representative gMAD pairs between VQA models induced by MDS-VQA and NoiseStability [13].

Figure 7. Representative challenging videos from YouTube-SFV SDR [37] selected by MDS-VQA. (a)-(f) show samples chosen without the diversity term, whereas (g)-(l) show samples chosen with diversity, resulting in a broader coverage of content and distortion patterns.