This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

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

Jian Zou[1], Xiaoyu Xu[1], Zhihua Wang[1], Yilin Wang[2], Balu Adsumilli[2], and Kede Ma[1*]
[1]City University of Hong Kong [2]Google Inc.
jian.zou@my.cityu.edu.hk, {yilin, badsumilli}@google.com,
{xiaoyxu, zhihua.wang, kede.ma}@cityu.edu.hk
https://github.com/Multimedia-Analytics-Laboratory/MDS-VQA

Abstract

Learning-based video quality assessment (VQA) has advanced rapidly, yet progress is increasingly constrained by a disconnect between model design and dataset curation. Model-centric approaches often iterate on fixed benchmarks, while data-centric efforts collect new human labels without systematically targeting the weaknesses of existing VQA models. Here, we describe MDS-VQA, a model-informed data selection mechanism for curating unlabeled videos that are both difficult for the base VQA model and diverse in content. Difficulty is estimated by a failure predictor trained with a ranking objective, and diversity is measured using deep semantic video features, with a greedy procedure balancing the two under a constrained labeling budget. Experiments across multiple VQA datasets and models demonstrate that MDS-VQA identifies diverse, challenging samples that are particularly informative for active fine-tuning. With only a 5% selected subset per target domain, the fine-tuned model improves mean SRCC from 0.651 to 0.722 and achieves the top gMAD rank, indicating strong adaptation and generalization.

1. Introduction

Video quality assessment (VQA) aims to predict perceptual video quality in a manner consistent with human judgments. Over the past two decades, VQA models have evolved from hand-crafted feature engineering tailored to specific distortion families [30, 39] to deep neural networks that learn perceptual cues directly from data [9, 41, 43], and more recently to large-scale pretrained models that leverage rich visual prior knowledge to achieve more human-aligned quality reasoning [44].

Despite this architectural progress, the field is increasingly constrained by a structural mismatch between *model-centric* innovation and *data-centric* curation. Model-centric VQA continuously refines architectures [36, 47], loss functions [4, 14], and training recipes [11, 12, 43], but typically does so on a small set of heavily reused benchmarks, cre-

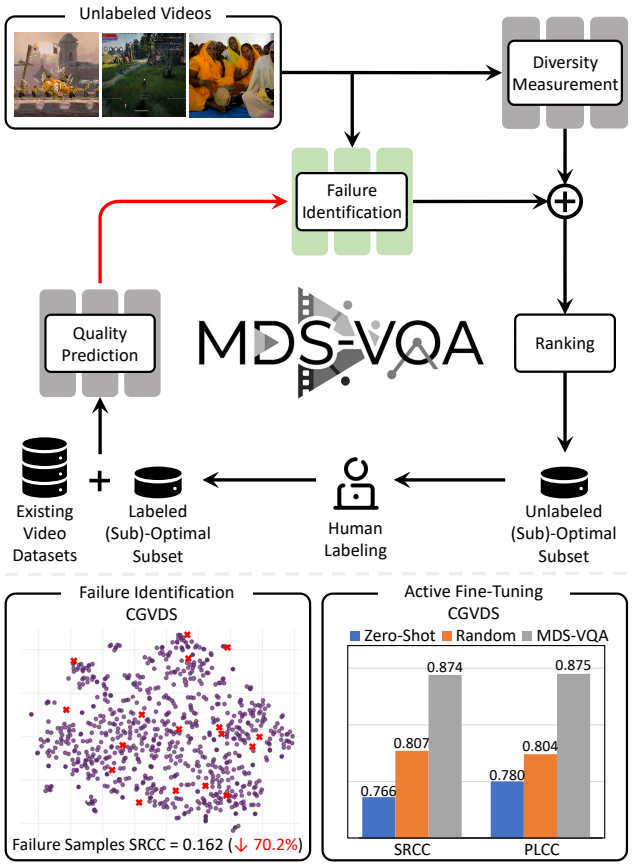

Figure 1. System diagram of MDS-VQA. We predict failure (*i.e.*, difficulty) on unlabeled target videos, combine difficulty and content diversity to select a small subset for human labeling, and actively fine-tune the VQA model on existing labeled data and the selected subsets. Bottom: on CGVDS [49], this 5% subset (marked by red crosses) yields strong failure identification and improved fine-tuning performance.

*Corresponding author.

ating persistent pressure to overfit dataset-specific idiosyncrasies. In parallel, data-centric VQA expends substantial resources on new subjective studies [25, 34, 35, 37, 49] to collect mean opinion scores (MOSs), yet these efforts often proceed without systematically targeting the failure modes of current top-performing VQA models [3, 38, 40].

A direct consequence of this inefficient loop is the *easy dataset* problem in VQA [28]. When datasets are dominated by content with easily identifiable distortions, even simple baselines (without spatiotemporal analysis and aggregation) can perform competitively, obscuring the limitations of advanced model architectures and reducing the marginal value of collecting more labels of the "same kind." As a result, new data collection—while valuable—does not always illuminate the true "blind spots" of contemporary VQA methods, and thus does not reliably drive improvements in cross-domain generalization.

In this work, we introduce MDS-VQA, a model-informed data selection mechanism for VQA that explicitly closes the feedback loop from models back to data (see Fig. 1). The key idea is to make dataset curation *model-aware*: rather than sampling videos for annotation passively or purely based on representativeness, we prioritize videos that are 1) *difficult* for the base model and 2) *diverse* in content. Concretely, we augment a base VQA model with an auxiliary failure predictor that estimates the difficulty of each unlabeled video (*i.e.*, how likely it is to expose the model's errors). We then combine these difficulty estimates with a content diversity measure to select a budgeted subset for human labeling. Active fine-tuning on the resulting "hard-and-diverse" subset completes the loop: model weaknesses guide data acquisition, and the acquired data directly improves the model.

Extensive experiments across multiple VQA datasets and models confirm that MDS-VQA is consistently more effective than competing data selection strategies at both 1) identifying failure samples and 2) improving quality prediction performance after active fine-tuning. In particular, using only a 5% selected subset per target domain, the fine-tuned model improves mean Spearman's rank correlation coefficient (SRCC) from 0.651 to 0.722, and achieves the top rank in a group maximum differentiation (gMAD) competition [17], supporting robust generalization beyond average-case correlation metrics.

2. Related Work

In this section, we briefly summarize prior work on VQA, covering both model-centric and data-centric advances.

2.1. Model-Centric VQA

The evolution of model-centric VQA reflects a continuous search for richer representations of perceptual quality. Early knowledge-driven methods [18] rely on hand-crafted features, often grounded in natural scene statistics, to capture statistical regularities in visual signals, but they struggle with the complex, mixed distortions typical of real-world videos. With deep learning, VQA moves from manual feature engineering to learned representations. Early deep VQA models [10] adopt 2D convolutional neural network (CNN) backbones, processing frames independently and aggregating temporal information afterward. To more directly encode spatiotemporal structure, 3D CNNs [31] are leveraged to directly learn from video volumes using 3D kernels [36], yet their limited receptive fields may hinder modeling long-range temporal dependencies. Transformer-based approaches [15] address this limitation by leveraging attention computation to capture global dependencies over long sequences, as exemplified by FAST-VQA [43].

More recently, vision-language models (VLMs) have been fine-tuned for VQA using quality scores, distributions, and/or descriptions, achieving improved performance. Subsequent work further applies reinforcement learning to optimize VLMs for VQA [44, 51], by aligning model outputs with MOSs to improve quality reasoning and scoring. While many reinforcement learning-based methods still retain a regression-style objective for quality prediction, VisualQuality-R1 [44] is a notable example that instead adopts a reinforcement-learning-to-rank formulation, which naturally handles varying MOS scales across datasets and supports efficient adaptation under limited supervision. For these advantages, we adopt VisualQuality-R1 to implement our base VQA model in the main experiments to push performance under a strong, state-of-the-art VLM setting.

2.2. Data-Centric VQA

Data-centric VQA is mainly driven by collecting human perceptual judgments of videos, mostly MOSs or difference MOSs. Such annotations require time-consuming subjective studies [6, 24], making large-scale, reliable datasets costly in both time and human effort. Early VQA datasets [24] are typically built from professionally generated content by applying synthetic distortions (*e.g.*, noise, blur, and compression artifacts) to a small set of pristine source videos. While crucial for early development, these datasets often lack diversity in both content and distortion, and models trained on them generalize poorly to real-world conditions [6, 20, 27, 47].

This limitation motivates a shift toward datasets with authentic distortions in user-generated content [7, 27, 35, 37]. Large-scale datasets like YouTube-UGC [35] and LSVQ [47] aim to dramatically increase diversity in content, scene complexity, capture conditions, and device characteristics, thereby encouraging VQA models to learn more perceptually relevant and generalizable representations. Recently, the proliferation of AI-generated videos has created a new data regime, where distortions are often semantic

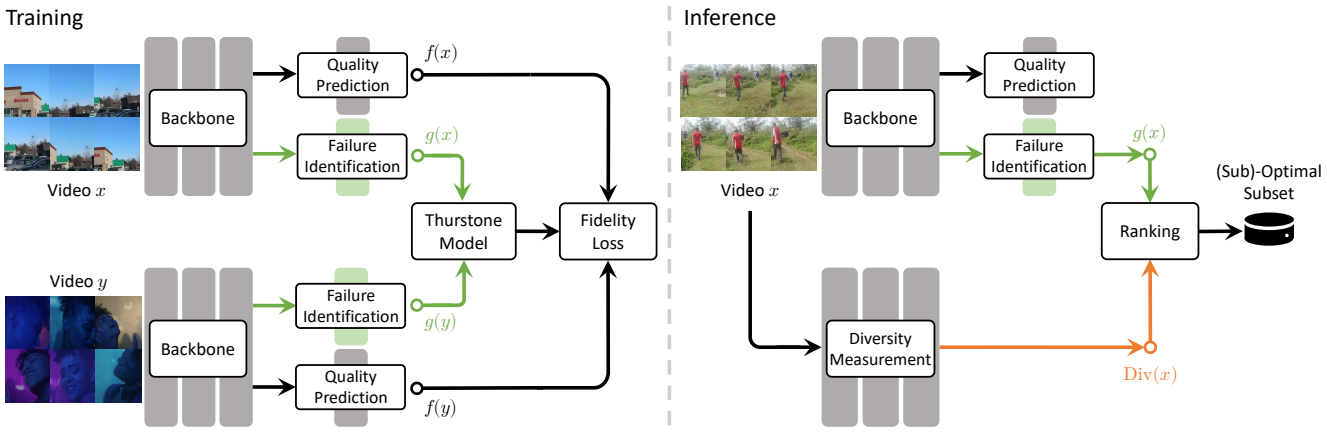

Figure 2. Training and inference of MDS-VQA. During training, we freeze the base quality model $f(\cdot)$ and optimize an auxiliary failure predictor $g(\cdot)$ by minimizing a fidelity loss under a Thurstone model [29]. During inference, we rank unlabeled videos by combining predicted difficulty scores with a content diversity measure to select a (sub-)optimal subset for human labeling and active fine-tuning.

or logical (*e.g.*, with unrealistic objects or unnatural motion) rather than purely signal-level artifacts. Datasets like AIGVQA-DB [34] have been introduced to target these emerging distortions. However, even with ongoing dataset expansion, a persistent issue remains: a large fraction of available data may still be easy for existing VQA models, which risks inefficient use of labeling resources and limits the impact of additional annotations [28, 41].

3. Proposed Method: MDS-VQA

In this section, we present MDS-VQA, a model-informed dataset selection mechanism that prioritizes informative unlabeled videos for annotation to improve VQA efficiency and generalization (see Fig. 2).

3.1. Overview and Problem Formulation

A productive development cycle for VQA should couple model improvement with data curation: models should be aware of where they fail, and new labels are then collected specifically to address those weaknesses. Motivated by this, we propose MDS-VQA, a model-informed data selection mechanism that prioritizes unlabeled videos most informative to the base VQA model.

Let $\mathcal{U}$ denote a large unlabeled video pool (*e.g.*, from a web crawl). Given a labeling budget, our goal is to select a subset $\mathcal{D} \subset \mathcal{U}$ for subjective annotation. MDS-VQA prioritizes videos that are both 1) difficult for the base VQA model and 2) diverse in content. Formally, we cast sample selection as a subset optimization problem [3]:

$$\mathcal{D} = \underset{\mathcal{S}\subset\mathcal{U}}{\operatorname{argmax}}\, \mathrm{Diff}(\mathcal{S}; f) + \lambda \mathrm{Div}(\mathcal{S}), \tag{1}$$

where $\mathcal{S}$ is a candidate video subset from the unlabeled pool $\mathcal{U}$ and $f(\cdot)$ is a top-performing VQA model. $\mathrm{Diff}(\cdot)$ measures how challenging the videos are for $f(\cdot)$, $\mathrm{Div}(\cdot)$ encourages content coverage, and λ trades off the two terms.

After annotating the selected subset $\mathcal{D}$, we incorporate the newly labeled data to retrain or fine-tune the base VQA model, thus closing the loop between model diagnosis and data acquisition.

Table 1. Structured text prompt used for training $g(\cdot)$.

You are doing a video quality assessment task. Here is the question: Assess how difficult it is to evaluate this video's quality. The difficulty rating should be a float between 1 and 5, rounded to two decimal places, with 1 representing very easy and 5 representing very difficult. Please only output the final answer with only one score in the `<answer> </answer>` tags.

3.2. Ranking-Based Difficulty Modeling

To instantiate $\mathrm{Diff}(\cdot)$, we augment the base VQA model $f(\cdot)$ with a failure predictor $g(\cdot)$ that estimates how likely a video is to expose the model's errors. In our main setting, $f(\cdot)$ is instantiated by VisualQuality-R1 [44] and $g(\cdot)$ is implemented by parameter-efficient low-rank adaptation (LoRA) [8], while keeping $f(\cdot)$ fixed.

LoRA-Based Failure Predictor. Concretely, we instantiate $g(\cdot)$ by attaching LoRA modules to the linear layers of the base model, while keeping the original weights fixed. For a linear projection with frozen weight $W_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, LoRA parameterizes the mapping as

$$W_{\text{LoRA}} = W_0 + \Delta W = W_0 + \frac{\alpha}{r} BA, \tag{2}$$

where $A \in \mathbb{R}^{r \times d_{\text{in}}}$ and $B \in \mathbb{R}^{d_{\text{out}} \times r}$ are trainable low-rank matrices, r is the rank, and α is a scaling factor. During training, we use a fixed structured prompt that asks the

model to output a difficulty rating in $[1, 5]$ (higher means harder), formatted in `<answer>` tags (see Table 1).

Ranking-Based Training Objective. An effective training objective is essential for reliably learning the failure predictor. Regressing absolute prediction error is sensitive to MOS scale differences across datasets, while binary "easy *vs*. hard" classification lacks granularity among difficult samples. We therefore formulate failure prediction as a learning-to-rank problem that only requires relative comparisons [3, 48]. Given a training pair of videos (x, y), the failure predictor outputs two scalars $g(x)$ and $g(y)$. Under a Thurstone model [29], we interpret these scores as the means of unit-variance Gaussian random variables, and compute the probability that video x is more difficult than y for the task of VQA as

$$\hat{p}(x, y) = \Phi\left(\frac{g(x) - g(y)}{\sqrt{2}}\right), \tag{3}$$

where $\Phi(\cdot)$ is the standard Gaussian cumulative distribution function. We construct supervisory signals from $f(\cdot)$'s predicted errors:

$$p(x, y) = \begin{cases} 1 & \text{if } |f(x) - \mu(x)| \geq |f(y) - \mu(y)|, \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where $\mu(x)$ and $\mu(y)$ denote the MOSs of x and y, respectively. We then optimize $g(\cdot)$ using the fidelity loss [32]:

$$\begin{aligned} \ell(x, y, p) = 1 &- \sqrt{p(x, y)\hat{p}(x, y)} \\ &- \sqrt{(1 - p(x, y))(1 - \hat{p}(x, y))}, \end{aligned} \tag{5}$$

which encourages $g(\cdot)$ to assign larger scores to videos that induce greater predicted errors in $f(\cdot)$.

3.3. Model-Informed Selection with Diversity

After training $g(\cdot)$, we combine its difficulty estimates with a content diversity measure to select a maximally informative subset $\mathcal{D} \subset \mathcal{U}$, which is then annotated and subsequently used for active fine-tuning.

Set Difficulty. For a candidate subset $\mathcal{S}$, we quantify its difficulty by the mean predicted failure score:

$$\mathrm{Diff}(\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} g(x), \tag{6}$$

where we omit the explicit dependence on $f(\cdot)$ in Eq. (1) for notation simplicity. A larger value indicates that the base VQA model struggles more on videos in $\mathcal{S}$, making them potentially more informative for improving generalization.

Set Diversity. To encourage broad coverage, we represent each video x as a set of frame-level semantic features $\mathcal{F}_x$, extracted using the CLIP vision encoder [22]. We then measure dissimilarity between two videos via the Chamfer distance between their frame-level features, capturing semantic variations beyond a single pooled descriptor:

$$\begin{aligned} d_{\mathrm{CD}}(\mathcal{F}_x, \mathcal{F}_y) = &\frac{1}{|\mathcal{F}_x|} \sum_{u \in \mathcal{F}_x} \min_{v \in \mathcal{F}_y} \|u - v\|_2^2 \\ &+ \frac{1}{|\mathcal{F}_y|} \sum_{v \in \mathcal{F}_y} \min_{u \in \mathcal{F}_x} \|v - u\|_2^2. \end{aligned} \tag{7}$$

The diversity of a subset $\mathcal{S}$ is then defined as the mean pairwise Chamfer distance:

$$\mathrm{Div}(\mathcal{S}) = \frac{1}{\binom{|\mathcal{S}|}{2}} \sum_{(x,y) \in \mathcal{S}} d_{\mathrm{CD}}(\mathcal{F}_x, \mathcal{F}_y). \tag{8}$$

Maximizing $\mathrm{Div}(\mathcal{S})$ encourages the selected videos to span a wide spectrum of scene types.

Greedy Approximation. Problem (1) is combinatorial and NP-hard [5, 19], so we adopt a greedy selection strategy. Starting from $\mathcal{D}_0 = \emptyset$, we iteratively add the video that best balances 1) its predicted difficulty and 2) its dissimilarity to already selected samples. At iteration $k + 1$, we choose

$$x^\star = \operatorname*{argmax}_{x \in \mathcal{U} \setminus \mathcal{D}_k} \left(g(x) + \frac{\lambda}{|\mathcal{D}_k|} \sum_{y \in \mathcal{D}_k} d_{\mathrm{CD}}(\mathcal{F}_x, \mathcal{F}_y) \right), \tag{9}$$

and update $\mathcal{D}_{k+1} = \mathcal{D}_k \cup \{x^\star\}$ until the labeling budget is reached. Here, λ is the same trade-off parameter.

By explicitly favoring samples that are both hard and non-redundant, MDS-VQA creates a targeted subset that more comprehensively reflects the base model's blind spots, leading to more efficient active fine-tuning.

3.4. Subset Labeling and Active Fine-Tuning

Subset Labeling. After obtaining the budgeted subset $\mathcal{D} \subset \mathcal{U}$ using Eq. (9), we conduct subjective experiments to acquire reliable human quality judgments for the selected videos using a standard protocol (*e.g.*, absolute category rating or paired comparison). Importantly, we represent the newly annotated data in a *scale-free* pairwise form by constructing labeled pairs from the annotated videos when needed. This pairwise formulation is invariant to the absolute scoring scale of any individual study, allowing the newly labeled subset to be merged seamlessly with existing pairwise VQA datasets without cross-dataset perceptual scale alignment [38, 40, 50].

Active Fine-Tuning with LoRA. We then update the base VQA model $f(\cdot)$ using both existing VQA data and the newly labeled subset. Our fine-tuning procedure follows the training recipe of VisualQuality-R1 (*i.e.*, prompting, on-the-fly pair construction within batches, and reinforcement-learning-to-rank optimization), but with one key modification: we replace full fine-tuning with parameter-efficient

Table 2. Failure identification performance. We compare MDS-VQA with eight competing methods on simulated unlabeled pools $\mathcal{U}$ from CGVDS, LIVE-Livestream, YouTube-SFV SDR, YouTube-SFV HDR2SDR, and AIGVQA-DB. Each entry reports SRCC/PLCC between the base model predictions and MOSs on the samples selected by each method; lower values indicate stronger failure identification. Best results are highlighted in bold.

Method	CGVDS [49]	LIVE-Livestream [25]	YouTube-SFV SDR [37]	YouTube-SFV HDR2SDR [37]	AIGVQA-DB [34]	Mean
Base model [44]	0.544 / 0.635	0.473 / 0.493	0.665 / 0.710	0.538 / 0.591	0.733 / 0.740	0.591 / 0.634
Random sampling	0.673 / 0.782	0.521 / 0.555	0.642 / 0.787	0.438 / 0.407	0.652 / 0.729	0.585 / 0.652
Core-set selection [23]	0.415 / 0.599	0.289 / 0.378	0.599 / 0.741	0.516 / 0.555	0.676 / 0.742	0.499 / 0.603
Sampling by RD [42]	0.411 / 0.412	0.253 / 0.352	0.495 / 0.535	0.372 / 0.592	0.640 / 0.660	0.434 / 0.510
MC dropout [21]	0.379 / 0.440	0.333 / 0.350	0.411 / 0.452	0.359 / 0.490	0.754 / 0.757	0.447 / 0.498
Greedy sampling [2]	0.331 / 0.564	0.461 / 0.586	0.472 / 0.532	0.334 / 0.451	0.659 / 0.663	0.451 / 0.559
ALCS [46]	0.388 / 0.817	0.310 / 0.545	0.485 / 0.696	0.342 / 0.614	0.597 / 0.600	0.424 / 0.654
FreeSel [45]	0.252 / 0.450	0.232 / 0.418	0.546 / 0.690	0.262 / 0.422	0.565 / 0.643	0.371 / 0.525
NoiseStability [13]	0.362 / 0.375	0.267 / 0.493	0.326 / 0.661	0.274 / 0.485	0.597 / 0.661	0.365 / 0.535
MDS-VQA (Ours)	**0.162 / 0.316**	**0.133 / 0.288**	**0.264 / 0.361**	**0.161 / 0.354**	**0.487 / 0.487**	**0.241 / 0.361**

LoRA, as stated in Sec. 3.2. This design mitigates overfitting or catastrophic forgetting while still facilitating efficient adaptation to model-challenging samples.

Iteration (Optional). The updated VQA model can be used to re-estimate difficulty and repeat Secs. 3.2-3.4 for additional rounds of selection when labeling budget permits, progressively shifting annotation effort toward the evolving failure modes of the VQA model.

4. Experiments

In this section, we assess MDS-VQA from three complementary angles: 1) failure identification (*i.e.*, how effectively MDS-VQA surfaces samples on which the base VQA model disagrees with human judgments), 2) active fine-tuning gains (*i.e.*, how much performance improves after fine-tuning on the selected subset), and 3) generalization under gMAD competition (*i.e.*, worst-case model differentiation beyond average-case correlation metrics).

4.1. Experimental Setups

Datasets. We use five VQA datasets that cover diverse video application scenarios: YouTube-UGC [35], CGVDS [49], LIVE-Livestream [25], YouTube-SFV+HDR [37], and AIGVQA-DB [34]. YouTube-UGC comprises approximately $1,500$ user-generated video clips with authentic distortions (*e.g.*, blockiness, blur, and jerkiness) and is used as the source dataset for training the base model. The other four datasets (in total over $42,000$ videos) serve as target-domain "unlabeled" pools. CGVDS [49] targets cloud-streamed gaming content encoded by hardware-accelerated H.264/MPEG-AVC at 60 frames per second under various bitrates and resolutions. LIVE-Livestream [25] focuses on 4K live streaming of professionally generated sports content, featuring characteristic impairments arising from capture artifacts and adverse network conditions. YouTube-SFV+HDR [37] contains short-form standard- and high-dynamic-range 1080p videos across popular categories (*e.g.*, dancing and cooking) with typical streaming distortions. AIGVQA-DB [34] benchmarks AI-generated videos produced by text-to-video models, capturing unique artifacts such as unrealistic objects and unnatural motion.

For worst-case evaluation, we additionally use LSVQ-1080p [47] (with $3,573$ high-resolution videos) to conduct gMAD testing, as it contains high-quality, subtle perceptual differences that are challenging for generalization.

Implementation Details. We compare the proposed MDS-VQA against a broad set of data selection strategies, including random sampling, core-set selection [23], sampling by representativeness diversity (RD) [42], Monte Carlo (MC) dropout [21], greedy sampling [2], ALCS [46], FreeSel [45], and NoiseStability [13].

We first train VisualQuality-R1 [44] on YouTube-UGC to obtain the base quality model. VisualQuality-R1 is built on Qwen2.5-VL [1], a video-pretrained backbone that supports temporal modeling through dynamic frame sampling and time-aligned embeddings, making it well suited for video-based quality reasoning and scoring. We then train the failure predictor $g(\cdot)$ by attaching LoRA modules to the base VQA model, using rank $r = 64$, scaling $\alpha = 128$, and dropout $p = 0.05$. Training is carried out by AdamW [16] with learning rate 1×10^{-5} followed by a linear decay schedule and batch size 8 for 10 epochs [26].

After learning $g(\cdot)$, we select informative samples from target-domain unlabeled pools using the greedy rule in Eq. (9) with $\lambda = 0.25$, under a fixed labeling budget of 5% per target dataset. Finally, we actively fine-tune the base model on the union of the full YouTube-UGC training set and the selected subsets from target domains. We use the same hyperparameter configuration as in training the failure predictor (see Supplementary for additional details).

Table 3. Active fine-tuning performance. Models are fine-tuned on the YouTube-UGC training set and subsets selected by different methods under a 5% labeling budget per target domain. Each entry reports SRCC/PLCC on the corresponding test set.

Method	YouTube-UGC [35]	CGVDS [49]	LIVE-Livestream [25]	YouTube-SFV SDR [37]	YouTube-SFV HDR2SDR [37]	AIGVQA-DB [34]	Mean
Base model [44]	0.708 / 0.709	0.766 / 0.780	0.561 / 0.587	0.666 / 0.718	0.495 / 0.557	0.711 / 0.748	0.651 / 0.683
Random sampling	0.760 / 0.756	0.807 / 0.804	0.569 / 0.628	0.703 / 0.761	**0.518** / 0.588	0.756 / 0.751	0.686 / 0.715
Core-set selection [23]	0.776 / 0.765	0.832 / 0.842	0.601 / 0.610	0.700 / 0.760	0.517 / 0.593	0.788 / 0.787	0.702 / 0.726
Sampling by RD [42]	0.795 / 0.784	0.780 / 0.825	0.597 / 0.602	0.729 / 0.778	0.517 / 0.587	0.772 / 0.767	0.698 / 0.724
MC dropout [21]	0.760 / 0.753	0.819 / 0.831	0.595 / 0.609	0.710 / 0.764	0.508 / 0.585	0.758 / 0.764	0.692 / 0.718
Greedy sampling [2]	0.781 / 0.773	0.831 / 0.834	0.585 / 0.605	0.711 / 0.774	0.500 / 0.580	0.775 / 0.777	0.697 / 0.724
ALCS [46]	0.785 / 0.772	0.807 / 0.823	0.626 / 0.636	0.727 / 0.781	0.513 / 0.579	0.789 / **0.792**	0.708 / 0.730
FreeSel [45]	0.814 / 0.798	0.832 / 0.849	0.627 / 0.646	0.719 / 0.787	0.498 / 0.590	0.789 / 0.785	0.713 / 0.742
NoiseStability [13]	0.806 / 0.798	0.826 / 0.847	0.601 / 0.616	0.727 / 0.780	0.500 / 0.594	**0.790** / 0.786	0.708 / 0.737
MDS-VQA (Ours)	**0.819 / 0.807**	**0.874 / 0.875**	**0.632 / 0.654**	**0.731 / 0.794**	0.507 / **0.595**	0.769 / 0.769	**0.722 / 0.749**

4.2. Main Results

Failure Identification Results. We first evaluate whether a selection method can reliably surface failure cases of the base VQA model. Concretely, after each method selects a 5% subset from a target-domain pool, we compute the SRCC and Pearson linear correlation coefficient (PLCC) between the base model predictions and MOSs on the selected subset. A lower correlation number indicates a more failure-focused subset as it reflects stronger disagreement with human judgments.

As shown in Table 2, across all evaluated target sets, MDS-VQA achieves the lowest (best) SRCC/PLCC results, indicating the strongest capability to identify difficult samples. For example, on CGVDS, MDS-VQA reduces SRCC to 0.162 compared with 0.673 for random sampling. The same pattern holds for live streaming, short-form streaming, and text-to-video generation, where MDS-VQA attains substantially lower SRCC values than alternatives. Notably, this advantage emerges even though neither the base VQA model nor the failure predictor sees target-domain data and labels. While the quality mapping learned by the base model may shift under domain changes, the resulting uncertainty and inconsistency patterns are more domain-agnostic and thus tend to transfer. This explains why MDS-VQA can better surface failure cases from target pools than either uncertainty-only [21] or diversity-only [2, 23, 42, 45, 46] selection.

Active Fine-Tuning Results. Next, we test whether the selected subsets are actually useful for improving the base VQA model. We fine-tune the model on the YouTube-UGC training set [35] and the selected 5% subsets from target domains, and evaluate performance across test splits of all datasets. From Table 3, we find that MDS-VQA yields the highest mean correlation, suggesting that the selected samples provide broad and transferable supervision rather than narrow domain-specific overfitting. We attribute these gains in part to explicitly enforcing *content diversity* during selection, which avoids repeatedly labeling near-duplicate hard cases and encourages coverage of complementary failure modes across varied content and distortion conditions (see also the results in Table 6).

Table 4. Quality prediction performance comparison under SRCC and gMAD competition [17]. We report each method's overall SRCC rank (based on the mean results in Table 3) and gMAD rank (evaluated on LSVQ-1080p [47]). Δ Rank denotes the difference between the two rankings.

Method	SRCC Rank	gMAD Rank	Δ Rank
MDS-VQA (Ours)	1	1	0
FreeSel [45]	2	2	0
NoiseStability [13]	3	6	-3
ALCS [46]	4	5	-1
Core-set selection [23]	5	8	-3
Sampling by RD [42]	6	3	3
Greedy sampling [2]	7	4	3
MC dropout [21]	8	9	-1
Random sampling	9	10	-1
Base model [44]	10	7	3

gMAD Competition Results. Average-case correlation can hide rare but consequential failures. We therefore conduct a gMAD competition [17] on LSVQ-1080p [47], which explicitly searches for samples that maximally distinguish two models and thus probes worst-case generalization. From Table 4, we observe that MDS-VQA attains the top gMAD rank (and also the top SRCC rank), indicating that its gains are consistent under both average-case and worst-case evaluation criteria. In contrast, several competing methods exhibit noticeable SRCC-gMAD rank discrepancies. The qualitative gMAD analysis in Fig. 3 further illustrates this generalization. When MDS-VQA is the attacker, it reveals that a core-set-induced model can severely under-score videos with animation and abstract patterns despite high MOSs, whereas the MDS-VQA-induced model remains more consistent with human perception when attacked by core-set selection [23].

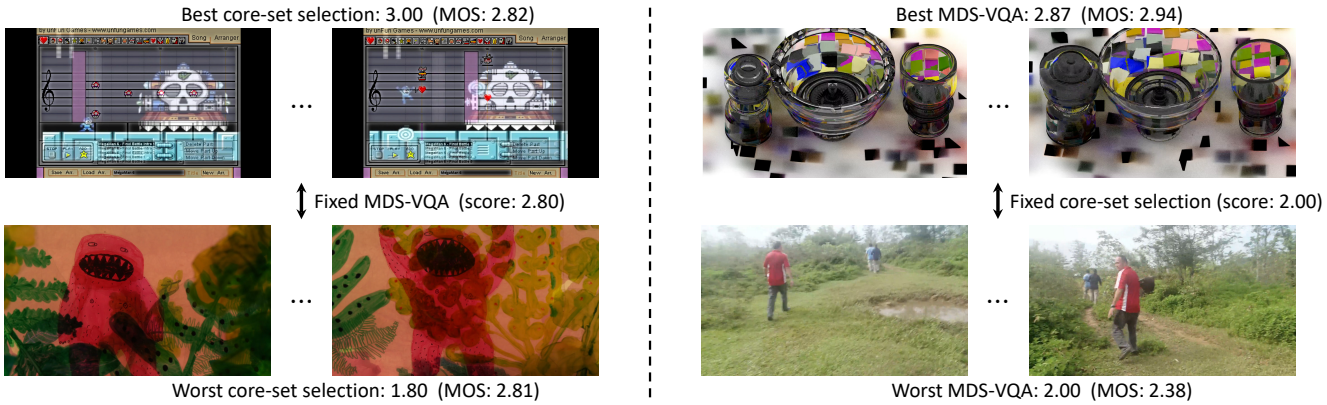

Figure 3. Representative gMAD pairs between VQA models induced by MDS-VQA and core-set selection [23]. Left: gMAD pairs found by fixing MDS-VQA-induced model predictions and searching for videos that maximally differentiate the core-set-induced model. Right: the reverse setting. Predicted scores and MOSs are shown for each pair, both on a $[1, 5]$ scale where higher values indicate better predicted and perceived quality, respectively.

Table 5. Ablation on training losses for the failure predictor $g(\cdot)$. The default setting is highlighted in bold.

Loss	Failure identification		Active fine-tuning	
	SRCC	PLCC	SRCC	PLCC
Base model [44]	0.665	0.710	0.666	0.718
Removing $g(\cdot)$	0.509	0.605	0.705	0.770
Classification loss	0.415	0.482	0.715	0.771
Regression loss	0.396	0.472	0.714	0.780
Ranking loss	**0.264**	**0.361**	**0.726**	**0.789**

Table 6. Ablation on the difficulty-diversity trade-off weight λ.

Weight	Failure identification		Active fine-tuning	
	SRCC	PLCC	SRCC	PLCC
Base model [44]	0.665	0.710	0.666	0.718
$\lambda = 0$	0.317	0.385	0.715	0.781
$\lambda = 0.125$	0.295	0.367	0.720	0.784
$\boldsymbol{\lambda = 0.250}$	**0.264**	0.361	**0.726**	**0.789**
$\lambda = 0.500$	0.281	**0.356**	0.725	0.784

4.3. Ablation Studies

We perform ablations to isolate the contributions of the key design components. Unless otherwise stated, results are reported on YouTube-SFV SDR [37], with base models trained on YouTube-UGC [35] under the default settings.

Effect of Failure Prediction Losses. We compare ranking-based training of the failure predictor $g(\cdot)$ against classification (with the cross-entropy loss) and regression (with the mean squared error), and also include a diversity-only variant that removes $g(\cdot)$ from selection. From Table 5, we see that the ranking loss consistently yields the strongest failure identification and the largest gains after active fine-tuning. This behavior is expected because our goal is inherently *ordinal*: we only need $g(\cdot)$ to induce a reliable relative ordering of samples by model difficulty, rather than to predict a calibrated absolute failure magnitude. In contrast, regression losses implicitly assume a stable numeric scale and can be dominated by a few high-error outliers or label noise. Classification further compresses supervision into coarse bins, discarding within-bin difficulty structure that is crucial for selecting the top fraction under a fixed budget. Additionally, ranking objectives are less sensitive to dataset-specific score ranges and heteroscedasticity, making them better aligned with cross-domain selection where the quality scale and error distribution may shift. Finally, the diversity-only variant underperforms because semantic coverage alone does not target the base model's blind spots, underscoring the importance of explicit difficulty modeling.

Effect of λ (Difficulty-Diversity Trade-Off). We vary λ in Eq. (9) to quantify the role of diversity. As shown in Table 6, introducing diversity steadily improves both failure identification and fine-tuning performance compared to $\lambda = 0$ (*i.e.*, difficulty-only), with the best overall balance at $\lambda = 0.25$. Increasing λ further to 0.5 slightly degrades performance, indicating that over-emphasizing coverage can dilute the "hardness" signal and reduce informativeness.

Effect of Diversity Features. We evaluate CLIP features [22] against those derived from SigLIP 2 [33] as well as internal representations extracted from the base VQA model, VisualQuality-R1 [44], for computing the diversity term. As detailed in Table 7, CLIP and SigLIP 2 yield comparable results, indicating that MDS-VQA is largely robust to the specific choice of diversity representations as long as they are generic visual embeddings trained at scale. In contrast, VisualQuality-R1 features perform the worst, likely because quality-oriented fine-tuning biases the representations toward perceptual degradations and reduces their

Table 7. Ablation on diversity feature representations.

Diversity feature	Failure identification SRCC	Failure identification PLCC	Active fine-tuning SRCC	Active fine-tuning PLCC
Base model [44]	0.665	0.710	0.666	0.718
VisualQuality-R1 [44]	0.397	0.452	0.720	0.783
CLIP [22]	**0.264**	0.361	0.726	0.789
SigLIP 2 [33]	0.279	**0.350**	**0.730**	**0.791**

Table 8. Ablation on the LoRA rank r for the failure predictor.

LoRA rank	Failure identification SRCC	Failure identification PLCC	Active fine-tuning SRCC	Active fine-tuning PLCC
$r = 8$	0.297	0.425	**0.727**	0.788
***r* = 64**	**0.264**	**0.361**	0.726	**0.789**
$r = 128$	0.273	0.418	0.724	**0.789**

ability to capture broad semantic variations for coverage. Given its widespread adoption and comparable performance to SigLIP 2, we use CLIP as a practical default.

Effect of LoRA Rank. We next ablate the LoRA rank r used to instantiate the failure predictor in Table 8. Setting a small rank $r = 8$ weakens failure identification, indicating insufficient capacity to learn reliable difficulty ordering. The default $r = 64$ achieves the best failure identification. Increasing to $r = 128$ degrades performance, suggesting mild overfitting or reduced cross-domain generalization with an over-parameterized adapter. Active fine-tuning, on the other hand, is largely stable across ranks, so we adopt $r = 64$ as the best accuracy-efficiency trade-off.

Generality across Base VQA Models. Finally, we apply MDS-VQA to two alternative VQA models with substantially different architectures: the CNN-based UVQ [36] and the Transformer-based ModularBVQA [41] using full fine-tuning. As shown in Table 9, in both cases, MDS-VQA achieves the best failure identification and active fine-tuning performance. These consistent gains across a reasoning-induced, prompt-based predictor and fully discriminative quality models suggest that the proposed "hard-and-diverse" selection principle underlying MDS-VQA is largely architecture-agnostic and can serve as a plug-and-play add-on for different VQA backbones.

5. Conclusion and Future Work

In this paper, we have presented MDS-VQA, a model-informed data selection mechanism that reconnects model-centric VQA development and data-centric dataset construction. MDS-VQA has been proven effective at identifying informative failure cases and facilitating active fine-tuning with a constrained labeling budget across multiple datasets and backbone architectures.

Looking forward, several promising and practical directions can further strengthen MDS-VQA. First, the current ranking-based failure predictor may yield *uncalibrated* difficulty scores, making the difficulty-diversity trade-off less convenient to tune. An adaptive, self-calibrating mechanism would substantially improve usability across deployment settings. Second, difficulty modeling can be enriched beyond a single scalar by incorporating fine-grained failure taxonomy (*e.g.*, spatial artifacts, temporal consistency, and semantic plausibility) so the selection process can target which weaknesses to fix. Third, diversity can be refined with spatiotemporal and motion-aware representations (and potentially multi-modal cues such as audio and metadata) to avoid selecting clips that are semantically different yet perceptually redundant, thereby improving coverage of distinct failure modes at the same labeling budget. Finally, a practical next step is to operationalize MDS-VQA in multi-round active learning with principled stopping criteria and cost-aware labeling (*e.g.*, mixing absolute category rating, paired comparison, and lightweight screening), enabling scalable dataset growth for emerging regimes such as AI-generated and streaming videos while maintaining annotation reliability and efficiency.

Table 9. Ablation on base VQA models.

Method	Failure identification SRCC	Failure identification PLCC	Active fine-tuning SRCC	Active fine-tuning PLCC
CNN-based UVQ [36]				
Base model [36]	0.590	0.644	0.577	0.648
Random sampling	0.577	0.632	0.591	0.664
Core-set selection [23]	0.417	0.389	0.625	0.707
Sampling by RD [42]	0.451	0.510	0.623	0.716
MC dropout [21]	0.278	0.374	0.636	0.719
Greedy sampling [2]	0.345	0.393	0.631	0.719
ALCS [46]	0.308	0.408	0.640	0.705
FreeSel [45]	0.462	0.525	0.645	0.723
NoiseStability [13]	0.224	0.305	0.661	0.720
MDS-VQA (Ours)	**0.101**	**0.137**	**0.677**	**0.729**
Transformer-based ModularBVQA [41]				
Base model [41]	0.636	0.732	0.573	0.706
Random sampling	0.606	0.696	0.659	0.743
Core-set selection [23]	0.269	0.718	0.686	0.769
Sampling by RD [42]	0.548	0.641	0.677	0.744
MC dropout [21]	0.414	0.454	0.680	0.758
Greedy sampling [2]	0.466	0.453	0.674	0.761
ALCS [46]	0.476	0.550	0.673	0.762
FreeSel [45]	0.412	0.711	0.689	0.769
NoiseStability [13]	0.342	0.402	0.694	0.769
MDS-VQA (Ours)	**0.124**	**0.384**	**0.713**	**0.771**

Acknowledgments

This work was supported in part by the Hong Kong ITC Innovation and Technology Fund (9440379 and 9440390), and a Google Gift Fund (9220141). We thank Tianhe Wu for assistance with the preparation of Figs. 1 and 2.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. 5

[2] Aditya Bhaskara, Sharvaree Vadgama, and Hong Xu. Greedy sampling for approximate clustering in the presence of outliers. In *Advances in Neural Information Processing Systems*, pages 11146–11155, 2019. 5, 6, 8

[3] Peibei Cao, Dingquan Li, and Kede Ma. Image quality assessment: Integrating model-centric and data-centric approaches. In *Conference on Parsimony and Learning*, pages 529–541, 2024. 2, 3, 4

[4] Pengfei Chen, Leida Li, Jinjian Wu, Weisheng Dong, and Guangming Shi. Unsupervised curriculum domain adaptation for no-reference video quality assessment. In *IEEE/CVF International Conference on Computer Vision*, pages 5158–5167, 2021. 1

[5] Geoff Davis, Stephane Mallat, and Marco Avellaneda. Adaptive greedy approximations. *Constructive Approximation*, 13 (1):57–98, 1997. 4

[6] Deepti Ghadiyaram, Janice Pan, Alan C. Bovik, Anush Krishna Moorthy, Prasanjit Panda, and Kai-Chieh Yang. In-capture mobile video distortions: A study of subjective behavior and objective algorithms. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(9):2061–2077, 2017. 2

[7] 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 *International Conference on Quality of Multimedia Experience*, pages 1–6, 2017. 2

[8] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 3

[9] Jari Korhonen, Yicheng Su, and Junyong You. Blind natural video quality prediction via statistical temporal features and deep spatial features. In *ACM International Conference on Multimedia*, pages 3311–3319, 2020. 1

[10] Patrick Le Callet, Christian Viard-Gaudin, and Dominique Barba. A convolutional neural network approach for objective video quality assessment. *IEEE Transactions on Neural Networks*, 17(5):1316–1327, 2006. 2

[11] 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 Transactions on Circuits and Systems for Video Technology*, 32(9):5944–5958, 2022. 1

[12] Dingquan Li, Tingting Jiang, and Ming Jiang. Unified quality assessment of in-the-wild videos with mixed datasets training. *International Journal of Computer Vision*, 129(4): 1238–1257, 2021. 1

[13] Xingjian Li, Pengkun Yang, Yangcheng Gu, Xueying Zhan, Tianyang Wang, Min Xu, and Chengzhong Xu. Deep active learning with noise stability. In *AAAI Conference on Artificial Intelligence*, pages 13655–13663, 2024. 5, 6, 8, 1, 2

[14] Wentao Liu, Zhengfang Duanmu, and Zhou Wang. End-to-end blind quality assessment of compressed videos using deep neural networks. In *ACM International Conference on Multimedia*, pages 546–554, 2018. 1

[15] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video Swin Transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022. 2

[16] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 5

[17] Kede Ma, Zhengfang Duanmu, Zhou Wang, Qingbo Wu, Wentao Liu, Hongwei Yong, Hongliang Li, and Lei Zhang. Group maximum differentiation competition: Model comparison with few samples. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):851–864, 2018. 2, 6

[18] Anish Mittal, Anush Krishna Moorthy, and Alan C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12): 4695–4708, 2012. 2

[19] Balas K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995. 4

[20] 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 Transactions on Image Processing*, 25(7):3073–3086, 2016. 2

[21] Remus Pop and Patric Fulop. Deep ensemble Bayesian active learning: Addressing the mode collapse issue in Monte Carlo dropout via ensembles. *arXiv preprint arXiv:1811.03897*, 2018. 5, 6, 8

[22] 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 *International Conference on Machine Learning*, pages 8748–8763, 2021. 4, 7, 8

[23] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018. 5, 6, 7, 8

[24] Kalpana Seshadrinathan, Rajiv Soundararajan, Alan C. Bovik, and Lawrence K. Cormack. Study of subjective and objective quality assessment of video. *IEEE Transactions on Image Processing*, 19(6):1427–1441, 2010. 2

[25] Zaixi Shang, Joshua P. Ebenezer, Alan C. Bovik, Yongjun Wu, Hai Wei, and Sriram Sethuraman. Assessment of subjective and objective quality of live streaming sports videos. In *Picture Coding Symposium*, pages 1–5, 2021. 2, 5, 6

[26] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. VLM-R1: A stable and generalizable R1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025. 5, 1

[27] Zeina Sinno and Alan C. Bovik. Large-scale study of perceptual video quality. *IEEE Transactions on Image Processing*, 28(2):612–627, 2018. 2

[28] Wei Sun, Wen Wen, Xiongkuo Min, Long Lan, Guangtao Zhai, and Kede Ma. Analysis of video quality datasets via design of minimalistic video quality models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(11): 7056–7071, 2024. 2, 3

[29] Louis L. Thurstone. A law of comparative judgment. *Psychological Review*, 34(4):273–286, 1927. 3, 4

[30] Hanghang Tong, Mingjing Li, Hongjiang Zhang, and Changshui Zhang. Blur detection for digital images using wavelet transform. In *IEEE International Conference on Multimedia and Expo*, pages 17–20, 2004. 1

[31] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3D convolutional networks. In *IEEE/CVF International Conference on Computer Vision*, pages 4489–4497, 2015. 2

[32] Ming-Feng Tsai, Tie-Yan Liu, Tao Qin, Hsin-Hsi Chen, and Wei-Ying Ma. FRank: A ranking method with fidelity loss. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 383–390, 2007. 4

[33] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad F. Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 7, 8

[34] Jiarui Wang, Huiyu Duan, Guangtao Zhai, Juntong Wang, and Xiongkuo Min. AIGV-Assessor: Benchmarking and evaluating the perceptual quality of text-to-video generation with LMM. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18869–18880, 2025. 2, 3, 5, 6

[35] Yilin Wang, Sasi Inguva, and Balu Adsumilli. YouTube UGC dataset for video compression research. In *IEEE International Workshop on Multimedia Signal Processing*, pages 1–5, 2019. 2, 5, 6, 7

[36] Yilin Wang, Junjie Ke, Hossein Talebi, Joong Gon Yim, Neil Birkbeck, Balu Adsumilli, Peyman Milanfar, and Feng Yang. Rich features for perceptual quality assessment of UGC videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13435–13444, 2021. 1, 2, 8

[37] Yilin Wang, Joong Gon Yim, Neil Birkbeck, and Balu Adsumilli. YouTube SFV+HDR quality dataset. In *IEEE International Conference on Image Processing*, pages 96–102, 2024. 2, 5, 6, 7, 1, 3

[38] Zhihua Wang and Kede Ma. Active fine-tuning from gMAD examples improves blind image quality assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4577–4590, 2021. 2, 4

[39] Zhou Wang, Alan C. Bovik, and Brian L. Evans. Blind measurement of blocking artifacts in images. In *IEEE International Conference on Image Processing*, pages 981–984, 2000. 1

[40] Zhihua Wang, Haotao Wang, Tianlong Chen, Zhangyang Wang, and Kede Ma. Troubleshooting blind image quality models in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16256–16265, 2021. 2, 4

[41] Wen Wen, Mu Li, Yabin Zhang, Yiting Liao, Junlin Li, Li Zhang, and Kede Ma. Modular blind video quality assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2763–2772, 2024. 1, 3, 8

[42] Dongrui Wu. Pool-based sequential active learning for regression. *IEEE Transactions on Neural Networks and Learning Systems*, 30(5):1348–1359, 2018. 5, 6, 8

[43] Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. FastVQA: Efficient end-to-end video quality assessment with fragment sampling. In *European Conference on Computer Vision*, pages 538–554, 2022. 1, 2

[44] Tianhe Wu, Jian Zou, Jie Liang, Lei Zhang, and Kede Ma. VisualQuality-R1: Reasoning-induced image quality assessment via reinforcement learning to rank. *arXiv preprint arXiv:2505.14460*, 2025. 1, 2, 3, 5, 6, 7, 8

[45] Yichen Xie, Mingyu Ding, Masayoshi Tomizuka, and Wei Zhan. Towards free data selection with general-purpose models. In *Advances in Neural Information Processing Systems*, pages 1309–1325, 2023. 5, 6, 8, 1, 2

[46] Xuyang Yan, Shabnam Nazmi, Biniam Gebru, Mohd Anwar, Abdollah Homaifar, Mrinmoy Sarkar, and Kishor Datta Gupta. A clustering-based active learning method to query informative and representative samples. *Applied Intelligence*, 52(11):13250–13267, 2022. 5, 6, 8, 1, 2

[47] Zhenqiang Ying, Maniratnam Mandal, Deepti Ghadiyaram, and Alan C. Bovik. Patch-VQ: 'Patching up' the video quality problem. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14019–14029, 2021. 1, 2, 5, 6

[48] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 93–102, 2019. 4

[49] Saman Zadtootaghaj, Steven Schmidt, Saeed S. Sabet, Sebastian Möller, and Carsten Griwodz. Quality estimation models for gaming video streaming services using perceptual video quality dimensions. In *ACM Multimedia Systems Conference*, pages 213–224, 2020. 1, 2, 5, 6

[50] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Transactions on Image Processing*, 30:3474–3486, 2021. 4

[51] Xuanyu Zhang, Weiqi Li, Shijie Zhao, Junlin Li, Li Zhang, and Jian Zhang. VQ-Insight: Teaching VLMs for AI-generated video quality understanding via progressive visual reinforcement learning. *arXiv preprint arXiv:2506.18564*, 2025. 2