Generalizable Video Quality Assessment via Weak-to-Strong Learning

Linhan Cao[1*], Wei Sun[2*♡], Xiangyang Zhu[3], Kaiwei Zhang[3], Jun Jia[1], Yicong Peng[1], Dandan Zhu[2], Guangtao Zhai[1], Xiongkuo Min[1†]
[1]Shanghai Jiao Tong University, [2]East China Normal University,
[3]Shanghai Artificial Intelligence Laboratory

Abstract

Video quality assessment (VQA) seeks to predict the perceptual quality of a video in alignment with human visual perception, serving as a fundamental tool for quantifying quality degradation across video processing workflows. The dominant VQA paradigm relies on supervised training with human-labeled datasets, which, despite substantial progress, still suffers from poor generalization to unseen video content. In this work, we explore ***weak-to-strong (W2S) learning*** *as a new paradigm for advancing VQA without reliance on human-labeled datasets. We first provide empirical evidence that a straightforward W2S strategy allows a strong student model to not only match its weak teacher on in-domain benchmarks but also surpass it on out-of-distribution (OOD) benchmarks, revealing a* ***distinct weak-to-strong effect in VQA****. Building on this insight, we propose a novel framework that enhances W2S learning from two aspects: (1)* ***integrating homogeneous and heterogeneous supervision signals*** *from diverse VQA teachers—including off-the-shelf VQA models and synthetic distortion simulators—via a learn-to-rank formulation, and (2)* ***iterative W2S training****, where each strong student is recycled as the teacher in subsequent cycles, progressively focusing on challenging cases. Extensive experiments show that our method achieves state-of-the-art results across both in-domain and OOD benchmarks, with especially strong gains in OOD scenarios. Our findings highlight W2S learning as a principled route to break annotation barriers and achieve scalable generalization in video quality assessment. Our data and code will be available at* `https://github.com/clh124/W2S-VQA`.

1. Introduction

Video quality assessment (VQA)[1] [33] plays an important role in modern video processing systems, delivering objective quality measurements used to optimize end-user Quality of Experience (QoE). With the advances in deep neural networks (DNNs) [7, 13, 28] and the increasing availability of human-annotated datasets [14, 45, 49, 54], current VQA models [46, 50–52] have achieved significant progress through supervised learning. Nevertheless, supervised learning inherently faces a limitation: **the generalization of the VQA models heavily depends on the diversity of the training data**. For example, even top-tier VQA models [46, 50–52] exhibit significant performance drops in out-of-distribution evaluations, as illustrated in Fig. 1.

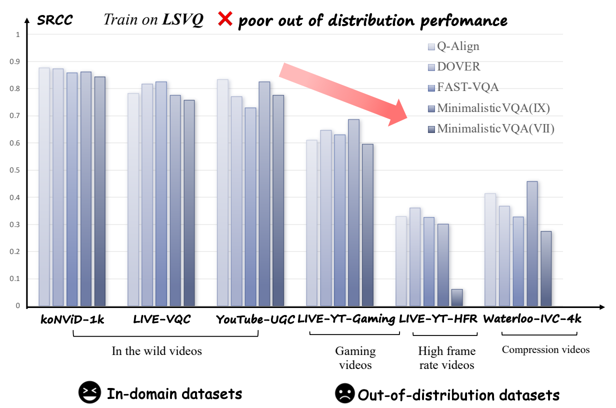

Figure 1. Significant performance drop of state-of-the-art models on out-of-distribution datasets.

Existing VQA research has primarily focused on constructing scene-specific datasets [22, 31, 44, 55] or large-scale datasets [10, 17] to improve model generalization across different video content and distortions. However, constructing such datasets is highly resource-intensive. A standardized subjective experiment comprises two key phases: **test sample curation** and **subjective quality annotation**. The test sample curation phase necessitates rigorous selection of representative video samples, as inadequate sampling strategies risk producing oversimplified datasets (*i.e.,* "*easy dataset*" problem [3, 46]) and may induce model overfitting. Meanwhile, subjective annotation—though vital—is laborious and costly. International Telecommunication Union (ITU) standards [15] outline specific recommendations for experimental setups, including display

[*]Corresponding author. ♡ Project Lead. † Corresponding author.
[1]This work focuses on no-reference (NR) or blind VQA, which assesses video quality without relying on additional reference information.

conditions, stimulus duration, subject count, and rating methodologies. These constraints, though necessary for statistically meaningful annotations, impede larger-scale dataset expansion due to prohibitive annotation costs.

Therefore, these limitations naturally raise an important question: *Can we train stronger VQA models without relying on large-scale human-annotated datasets?* Prior efforts have investigated self-supervised and unsupervised VQA approaches [4–6, 32, 34] which primarily employ contrastive learning with proxy tasks such as distortion-type or severity classification on synthetically generated data. However, these approaches struggle to capture the complex and nonlinear degradation patterns present in real-world videos, limiting their ability to model authentic distortions. As a result, their performance still lags significantly behind supervised counterparts on in-the-wild VQA datasets.

Recent progress in **weak-to-strong (W2S) generalization** [1, 11] provides a promising approach for tackling this open problem. In this paradigm, a strong student model—equipped with higher learning capacity or powerful pre-trained knowledge—can learn effectively from the supervision of a weaker model and further generalize to hard examples beyond the teacher's reach. It is thus natural to leverage an existing VQA model as a weak teacher to distill a stronger one, obviating the need for human-annotated labels. This approach raises two critical questions: (1) *How effectively does W2S generalization apply to VQA, a task that inherently involves subjective human perception rather than deterministic high-level semantics*, and (2) *How can we enhance its performance to meet the demands of practical VQA applications?*

This work investigates these two problems. First, we empirically demonstrate that a straightforward W2S generalization approach enables the student model to match the performance of its weak teacher (*e.g.,*, off-the-shelf VQA models) on in-domain benchmarks and surpass it on out-of-domain (OOD) benchmarks, revealing a clear weak-to-strong generalization effect in VQA.

Second, we advance W2S learning for VQA from two aspects: **integrating diverse supervision signals** and **iterative W2S training**. For the former, we incorporate multiple types of "VQA models" as weak models to refine and diversify the supervised signals, including (1) ensembling homogeneous VQA models (*i.e.,* off-the-shelf VQA models) to improve the reliability of supervision, and (2) integrating heterogeneous teachers (*i.e.,* synthetic distortion simulators) to enrich the supervision space. To unify these heterogeneous supervision signals, we reformulate quality regression as a **ranking problem** to make the model to learn quality assessment capabilities through pairwise comparisons. For the latter, we propose an iterative W2S learning strategy with difficulty-guided sampling, where each trained strong model is recycled as the weak teacher for the next iteration. Within each cycle, we deliberately select difficult samples so that subsequent models focus on challenging cases beyond the reach of weaker teachers, thereby progressively expanding the generalization capacity of the student model.

Our key contributions are summarized as follows:

- We empirically validate a distinct W2S generalization effect in VQA, providing a new paradigm for advancing self-supervised and weakly supervised approaches for VQA.
- We introduce a novel W2S generalization framework that integrates heterogeneous supervision signals from diverse teachers and incorporates an iterative W2S training strategy.
- Within this framework, our student model achieves state-of-the-art results on both in-domain and OOD benchmarks, with particularly notable gains on OOD performance.

2. Related Work

2.1. VQA Models

Supervised VQA. Early VQA models [36, 41] were largely knowledge-driven, extracting handcrafted features (*e.g.*, natural scene statistics [35], motion cues [18]) to quantify distortions and training shallow regressors for quality prediction. Subsequent approaches [21, 54] shifted to representation learning, employing pre-trained DNNs to extract frame-level quality representations, coupled with sequence models such as GRUs or Transformers for temporal regression. More recent efforts adopt end-to-end fine-tuning of advanced vision architectures, including Vision Transformers (ViTs) [7] and large multimodal models (LMMs) [52], with the designs such as grid-based mini-patch sampling or key-frame selection to mitigate the computational burden of full-video training. While these advancements have significantly improved the performance of VQA models on in-domain datasets, they still struggle to generalize satisfactorily to OOD datasets.

Weakly-supervised VQA. Existing weakly supervised VQA methods typically adopt full-reference (FR) quality assessment models as pseudo-supervision, either relying on a single teacher [24, 57] or combining multiple teachers through multi-task [23] or ensemble-learning frameworks [27, 53]. However, these approaches primarily target representation learning, and their performance is generally inferior to that of the teacher FR models; they also often require additional fine-tuning on human-labeled datasets to remain competitive. Moreover, they are mainly designed for synthetic distortions, making them unsuitable for in-the-wild videos where no pristine reference exists. In contrast, our study shows that even a single weak teacher can offer sufficiently informative supervision to train a strong student model that surpasses the teacher itself, while remaining suitable for no-reference quality assessment settings.

VQA as Ranking. Ranking-based methods reformulate quality prediction from a regression problem into a ranking problem. To this end, various loss functions such as hinge loss [25], fidelity loss [56], binary cross-entropy loss [58], and differentiable approximations of Spearman Rank Correlation loss [19] have been employed to learn relative quality rankings from pairwise comparisons or groups of samples. Such methods are particularly effective in mitigating the misalignment of quality scales across different datasets and can be applied in scenarios where only relative quality labels are available. Consequently, they have been widely adopted in weakly supervised training and mixed-dataset training. In this work, we also adopt a learning-to-rank strategy to unify the heterogeneous supervisory signals provided by diverse weak teachers.

2.2. Weak-to-strong Generalization

Weak-to-strong (W2S) generalization studies how strong models can learn from weaker supervision yet surpass their teachers. Early empirical studies [1] showed that simply fine-tuning a strong model on weak labels already allows the student to outperform its weak teacher across domains such as NLP, reward modeling, and games. Building on these foundations, subsequent studies have focused on improving the quality of weak supervision. Co-supervised and mixture-of-experts approaches [26] combine diverse weak teachers to mitigate noise and bias; ensemble and scalable oversight methods [43] enhance teacher reliability through aggregation and debate mechanisms; and confidence-aware objectives [1, 11] further balance weak guidance with student predictions to avoid overfitting to noisy labels. Inspired by these advancements, we leverage diverse weak teachers to diversify and improve the supervision signals.

3. Weak-to-Strong Learning for VQA

3.1. Problem Setup

Assume that we have access to a weak VQA model f_{weak}, which in practice can be instantiated by existing open-source VQA models. Let $D_{\text{w2s}} = \{x_1, x_2, \ldots, x_n\}$ denote an unlabeled video dataset with no ground-truth labels. We use f_{weak} to generate predictions $\hat{y}_j = f_{\text{weak}}(x_j)$ for each video $x_j \in D_{\text{w2s}}$, and subsequently train or fine-tune a strong student model f_{w2s} on D_{w2s} using these predictions as supervision. The objective is to examine whether f_{w2s} can outperform f_{weak} without relying on human annotations for training.

3.2. Weak-to-Strong Implementation for VQA

Weak Models f_{weak}. We select five open-source VQA models[2] f_{weak}: MinimalisticVQA (VII) [46], MinimalisticVQA (IX) [46], FAST-VQA [50], DOVER [51], and Q-Align [52]. All models are trained on the LSVQ dataset [54] and encompass architectures including convolutional neural networks, vision transformers, and LMMs. Detailed descriptions of these methods and the rationale behind their selection are provided in the *Supp. Sec.* B.1.

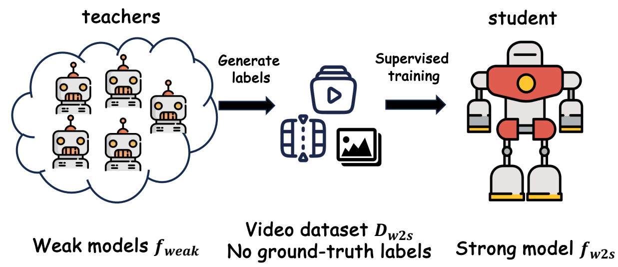

Figure 2. Overview of our weak-to-strong training pipeline.

Strong Model f_{w2s}. For the strong student model, we adopt an LMM backbone with substantially higher capacity than the weak teachers, using `LLaVA-OneVision-Chat-7B` [20] as a representative example. A detailed parameter and architecture comparison between weak and strong models is provided in Supp. Table 4. We additionally evaluate several other state-of-the-art LMMs as strong students, with their results summarized in *Supp. Sec.* D.2. The strong model reported in the main paper corresponds to the best-performing backbone among all candidates. To better adapt it to the VQA task, we follow a preprocessing strategy similar to LMM-VQA [9]: one key frame per second is sampled for the vision encoder, while motion features are extracted for each key frame using all frames within that second via SlowFast [8]. These motion features are then processed by a motion projector and fused with the visual features before being fed into the language model of the LMM. A detailed description of our student model is provided in *Supp. Sec.* C.1, and its overall architecture is illustrated in Figure 3.

Training Dataset D_{w2s}. We first collect a pool of 3 million videos from popular social media platforms, including YouTube, TikTok, Youku, and Bilibili. From this pool, we select a subset based on nine low-level metrics that quantify visual characteristics—blockiness [40], blur [37], contrast [39], noise, flickering [38], colorfulness [12], luminance, temporal information, and spatial information [15]—to ensure that the selected videos are as diverse as possible across these dimensions. We then sample 200k videos from the matched subset to construct a representative and diverse training set for the student model, covering a wide range of quality conditions. A detailed description of the dataset construction procedure and analysis is provided in *Supp. Sec.* A.

Training Protocol. We train f_{w2s} on D_{w2s}, where supervi-

[2] In this context, the term "weak" refers to their capability relative to the student model. In fact, the selected models represent state-of-the-art VQA approaches.

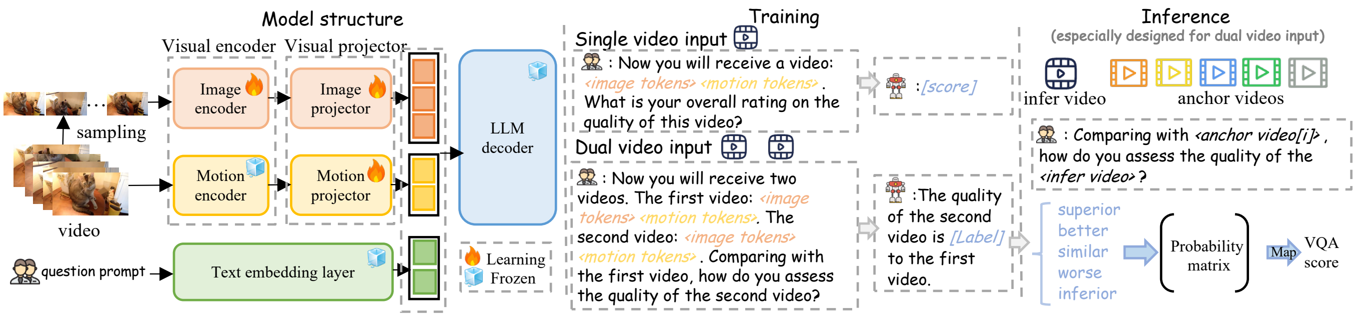

Figure 3. Overall architecture of our strong student model. Following LMM-VQA [9], we use a dual-branch visual encoder with an additional motion module for temporal distortion modeling. The model supports both single- and dual-video input strategies with distinct training and inference designs. For single-video input, the model directly predicts the quality score. For dual-video input, it is trained to predict relative quality between two videos and converts it into an absolute score through a designed inference strategy.

sion is provided by pseudo-labels generated from f_{weak}, and optimize the model with the standard cross-entropy loss. Training is conducted with AdamW, an initial learning rate of 1×10^{-4}, a cosine decay schedule, and a weight decay of 0.05. We use a batch size of 8 and train for 25k iterations with linear warm-up in the first 750 steps. All experiments are implemented in PyTorch and trained on 8 NVIDIA H200 GPUs. We intentionally adopt **standard experiment settings** (*e.g.*, model architectures) consistent with prior work to ensure that *the observed performance gains stem from the W2S framework itself*, rather than architectural or other modifications.

Validation Datasets. To comprehensively assess model performance, we evaluate on ten VQA benchmarks grouped into *in-domain* and *out-of-distribution* (OOD) categories. The in-domain datasets include LSVQ Test [54], LSVQ 1080p [54], KoNViD-1k [14], LIVE-VQC [45], and YouTube-UGC [49], all consisting of user-generated content (UGC) videos. The OOD datasets comprise LIVE-YT-Gaming [55], CGVDS [42], LIVE-YT-HFR [31], Waterloo-IVC-4K [22], and KVQ [29], which differ from in-domain benchmarks in both content distribution and distortion types. Further details of these datasets are provided in *Supp. Sec.* A.4.

Evaluation Metrics. We adopt two widely used criteria to evaluate the performance of VQA models: Spearman Rank Correlation (SRCC) and Pearson Linear Correlation (PLCC), which indicate the prediction monotonicity and prediction linearity, respectively.

3.3. Experimental Results and Analysis

We report overall in-domain and OOD performance in Table 1. For student models supervised by weak teachers, we present results trained on (i) a randomly sampled subset of our data with the same scale as LSVQ (27k videos), and (ii) the full large training set (200k videos). When training on 27k videos, for in-domain benchmarks, the student model achieves performance comparable to its teachers, with only a minor degradation of 0.15%, While for OOD benchmarks, the student exhibits substantial average gains of 6.05% over its teachers, highlighting a pronounced weak-to-strong generalization effect. Interestingly, for stronger teacher models such as MinimalisticVQA (IX) and Q-Align, we observe that their student counterparts achieve comparable performance on in-domain benchmarks and even surpass the supervised models on OOD benchmarks. When further scaling up the training data to 200k videos, we observe consistent performance gains on both in-domain and OOD benchmarks, reflecting the benefits of increased data diversity and broader visual coverage. Importantly, such scaling is easily supported by our W2S training pipeline, whereas achieving comparable expansion under human-labeled supervision is considerably more expensive and challenging.

In summary, our results empirically demonstrate a clear

Table 1. Performance comparison of weak teachers, students trained with weak teacher labels at two data scales (27k vs. 200k), and students trained with LSVQ ground-truth labels. Best performance in each category is indicated in **bold**.

Methods	In-domain		OOD	
	SRCC	PLCC	SRCC	PLCC
Weak Teachers				
MinimalisticVQA(VII)	0.817	0.830	0.490	0.551
MinimalisticVQA(IX)	**0.849**	0.859	**0.574**	**0.622**
FAST-VQA	0.838	0.849	0.486	0.512
DOVER	0.842	0.845	0.519	0.569
Q-Align	0.844	**0.861**	0.555	0.606
Students Supervised by Weak Teacher Labels (27k)				
MinimalisticVQA(VII)-labeled	0.818	0.831	0.515	0.576
MinimalisticVQA(IX)-labeled	0.845	0.853	**0.582**	**0.633**
FAST-VQA-labeled	0.836	0.846	0.543	0.597
DOVER-labeled	0.840	0.849	0.549	0.611
Q-Align-labeled	**0.846**	**0.857**	0.578	0.632
Students Supervised by Weak Teacher Labels (200k)				
MinimalisticVQA(VII)-labeled	0.824	0.833	0.536	0.593
MinimalisticVQA(IX)-labeled	**0.849**	**0.859**	**0.591**	**0.639**
FAST-VQA-labeled	0.842	0.845	0.550	0.604
DOVER-labeled	0.843	0.850	0.554	0.617
Q-Align-labeled	0.848	**0.859**	0.581	0.636
Students Supervised by Ground Truth Labels (27k)				
LSVQ-labeled	**0.851**	**0.861**	**0.577**	**0.608**

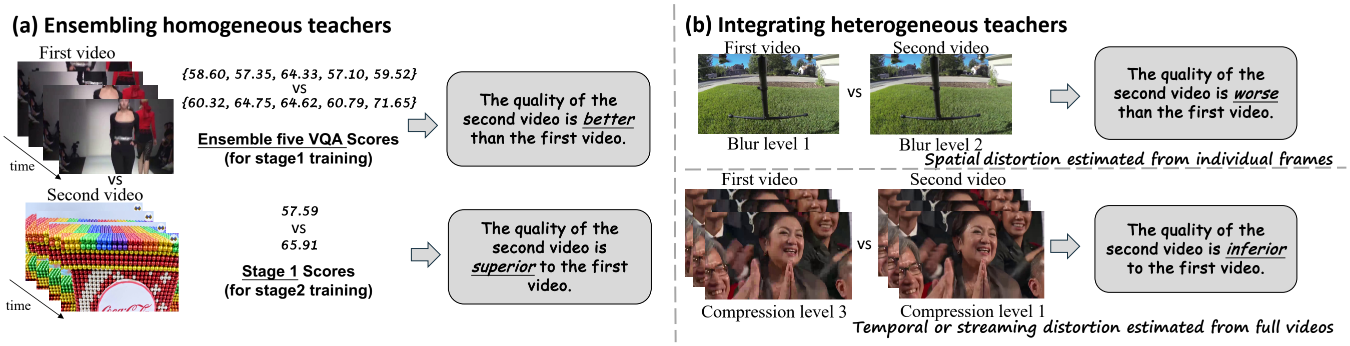

Figure 4. Our pairwise quality annotations consist of two types: (1) pseudo-labeling based on ensembling homogeneous teachers, and (2) quality ranking derived from integrating heterogeneous teachers.

weak-to-strong generalization effect in VQA, where the most significant improvements arise on OOD data unseen during training. This finding is particularly important for VQA, as in-domain performance on existing benchmarks has largely saturated and even risks overfitting, while current methods suffer from severe degradation on OOD scenarios. Weak-to-strong generalization therefore offers a promising paradigm for addressing this challenge, and in the next section we present a practical solution.

4. Improving Weak-to-Strong Learning for VQA

We enhance weak-to-strong generalization in VQA from two aspects: (1) unifying diverse supervision signals and (2) iterative W2S training, both aimed at expanding the generalization capacity of the student model.

4.1. Unifying Diverse Supervision Signals

4.1.1. Ranking-based VQA Method

Absolute quality scores obtained from different labeling manners may be inconsistent in their ranges and scales, making them unsuitable for regression-based training. In contrast, the relative quality ranks of video pairs within the same manner are consistent. To unify these heterogeneous supervision signals, we reformulate quality prediction as a **ranking problem**, enabling the model to learn quality assessment capability through pairwise comparisons.

Specifically, given a video pair $(\boldsymbol{x}^A, \boldsymbol{x}^B)$, we input them into the student model defined in Section 3.2, which is trained to predict their relative quality. Following [58], we adopt ranking labels {"superior", "better", "similar", "worse", "inferior"} to refine ranking accuracy. During inference, we employ the adaptive soft comparison method [58] to derive quality scores. It first computes a soft probability matrix over ranking categories by comparing each test video against anchor videos, and then applies maximum a posteriori (MAP) estimation [48] under Thurstone's Case V model [47] to obtain calibrated quality scores. The detailed inference procedure is provided in *Supp. Sec.* C.3.

4.1.2. Ensembling Homogeneous Teachers

In Section 3.3, we observe that stronger teacher models generally yield more capable students, in some cases even surpassing fully supervised counterparts. A naïve strategy is thus to enhance the accuracy of teacher models. To this end, we adopt a simple approach: averaging ensemble predictions from five VQA methods in Section 3.2 to improve the reliability of the supervision signals.

For video pair generation, given a pair (x^A, x^B), each VQA model $f_{\text{weak},i}$ produces quality scores $\hat{y}_i^A$ and $\hat{y}_i^B$. We compute the mean scores $\overline{y}^A$ and $\overline{y}^B$[3], and the score variances σ_A^2 and σ_B^2. Assuming the quality difference $\Delta = \overline{y}^A - \overline{y}^B$ follows a Gaussian distribution $\mathcal{N}(\Delta; 0, \sigma_\Delta^2)$ with $\sigma_\Delta = \sqrt{\sigma_A^2 + \sigma_B^2}$, labels are assigned according to the statistical significance thresholds in [58]: "superior" if $\Delta > 2\sigma_\Delta$, "better" if $\sigma_\Delta < \Delta \leq 2\sigma_\Delta$, "similar" if $-\sigma_\Delta < \Delta \leq \sigma_\Delta$, "worse" if $-2\sigma_\Delta < \Delta \leq -\sigma_\Delta$, and "inferior" if $\Delta \leq -2\sigma_\Delta$.

4.1.3. Integrating Heterogeneous Teachers

Another complementary approach is to diversify the teacher models in order to enrich the supervision signals. In this work, we leverage **synthetic distortion simulators** as specialized VQA models, which do not require human annotations for training and can be easily scaled. Concretely, we introduce three categories of synthetic distortions to emulate typical real-world degradations: **spatial distortions**, **temporal distortions**, and **streaming distortions**. Spatial distortions include *resolution downscaling, Gaussian blur, Gaussian noise, darkening*, and *brightening*, simulating capture-related artifacts. Temporal distortions cover *jitter* and *stuttering*, which mimic playback issues often observed in practice. Streaming distortions involve *H.264* and

[3]We use a four-parameter logistic function to map the predicted scores from different weak models onto a common scale for fair comparison and subsequent evaluation.

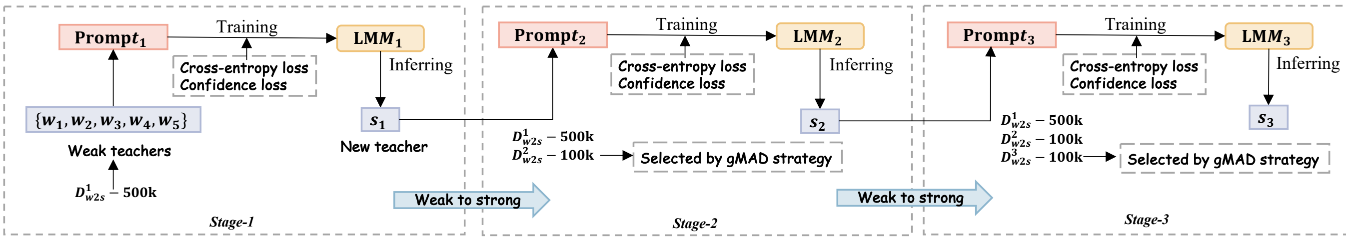

Figure 5. The framework of our iterative weak-to-strong training strategy.

H.265 compression, capturing compression artifacts introduced by modern media delivery platforms. The detailed simulation procedures are provided in *Supp. Sec.* A.3.

We leverage distortion severity levels (*e.g.*, constant rate factor for compression) as pseudo-labels to infer relative quality. Given a primary video x^0 and a synthetic distortion simulator $\mathcal{S}$, we degrade x^0 across $N_{\mathcal{S}}$ severity levels to generate distorted videos $\{x^i_{\mathcal{S}}\}_{i=1}^{N_{\mathcal{S}}}$. Pairs $(x^i_{\mathcal{S}}, x^j_{\mathcal{S}})$ are randomly sampled. Pairs with a severity difference $|i-j|>1$ are labeled as "superior" or "inferior" depending on the relative order of i and j, while pairs with $|i-j|=1$ receive "better" or "worse". The "similar" label is intentionally excluded, as $i-j=0$ implies identical videos.

4.2. Iterative Weak-to-Strong Training Strategy

Within our W2S training framework, we have demonstrated that the student model can surpass its teacher models. This observation naturally motivates an iterative strategy: *once a student model is trained, it can be promoted to act as a new teacher, thereby enabling another round of weak-to-strong training*. Through such iterative cycles, the student progressively inherits knowledge from its predecessors while further enhancing its generalization capability. Therefore, we adopt this iterative paradigm to continually refine the student model.

From the data perspective, we expect the training samples in the next iteration to pose challenges beyond the capacity of the current teacher models, thereby further expanding the capability of the student. To this end, we introduce a difficult-sample selection strategy for both types of supervision signals in Section 4.1. Specifically, given a student model $f^{(i)}_{\text{w2s}}$ trained in the i-th iteration, the construction of difficult samples is straightforward for synthetic distortion pairs described in Section 4.1.2, since ground-truth labels can be directly derived from the distortion levels. We use $f^{(i)}_{\text{w2s}}$ to infer the relative quality of these pairs and select only those misclassified by the student as the training data for the $(i+1)$-th iteration.

While for the video pairs described in Section 4.1.2, no ground-truth labels are available. To address this, we adopt the group maximum differentiation (gMAD) competition framework [30] to select pairs that exhibit the largest disagreement between VQA models. Given the weak model set $\{f^j_{\text{weak}}\}_{j=1}^{N_{\text{weak}}}$ used to train $f^{(i)}_{\text{w2s}}$, we first partition the video pool $D^{(i+1)}_{\text{w2s}}$ into ξ uniform quality levels based on the predictions of f^j_{weak}, within which videos are assumed to have similar perceptual quality. We then select pairs that are maximally differentiated by the trained student model $f^{(i)}_{\text{w2s}}$ while indistinguishable to the weak model f^j_{weak} by

$$
\begin{aligned}
(\hat{x}^A, \hat{x}^B) \in \arg \max_{x^A,\, x^B \in D^{(i+1)}_{\text{w2s}}} & \left[f^{(i)}_{\text{w2s}}(x^A) - f^{(i)}_{\text{w2s}}(x^B)\right] \\
\text{s.t.} \quad & \left|f^j_{\text{weak}}(x^A) - f^j_{\text{weak}}(x^B)\right| \leq \xi.
\end{aligned}
\tag{1}
$$

Moreover, we also reverse the roles of f^j_{weak} and $f^{(i)}_{\text{w2s}}$ to capture cases where the student perceives similar quality but the weak model disagrees. This strategy systematically exploits the decision boundary mismatches between student and teacher models, generating informative and challenging samples that drive further improvements in next-round W2S training.

4.3. Training Strategy

We employ the standard cross-entropy loss as a baseline objective. However, weak annotations inevitably contain noise, and directly supervising the student with cross-entropy risks overfitting to erroneous labels. To mitigate this, we introduce an auxiliary confidence loss [1, 11] that encourages the student to reinforce its own confident predictions, particularly when they diverge from weak labels. The overall objective is formulated as

$$
\mathcal{L} = (1-\lambda)\,\mathcal{L}_{\text{CE}} + \lambda\,\mathcal{L}_{\text{conf}}, \tag{2}
$$

where $\mathcal{L}_{\text{CE}}$ denotes the cross-entropy loss, $\mathcal{L}_{\text{conf}}$ the confidence loss, and λ adaptively balances label reliability against model predictions. Details of the confidence loss are provided in *Supp. Sec.* C.2.2.

For training data, we construct a total of 700k annotated video pairs using the procedure described in Section 4.1.2 and Section 4.1.3. These pairs are partitioned into three subsets of 500k, 100k, and 100k, denoted as $D^{(1)}_{\text{w2s}}$, $D^{(2)}_{\text{w2s}}$, and $D^{(3)}_{\text{w2s}}$, corresponding to the three stages of iterative training. A detailed breakdown of the dataset, as well as the complete training setup, is provided in *Supp. Sec.* A.1 and *Supp. Sec.* C.2.1.

Table 2. Performance comparison with competing methods. The single-teacher supervision baseline is defined by the best-performing model reported in Table 1. The best and second-best results are marked by **red** and blue. "Overall" represents the weighted average results based on the number of videos in each dataset.

In-domain Datasets	LSVQ$_{test}$		LSVQ$_{1080p}$		KoNViD-1k		LIVE-VQC		YouTube-UGC		Overall	
# of videos	7,182		3,573		1,200		585		1,020		-	
Methods	**SRCC**	**PLCC**	**SRCC**	**PLCC**	**SRCC**	**PLCC**	**SRCC**	**PLCC**	**SRCC**	**PLCC**	**SRCC**	**PLCC**
Competing Methods: Teachers												
MinimalisticVQA(VII) [46]	0.861	0.859	0.740	0.784	0.843	0.841	0.757	0.813	0.775	0.779	0.817	0.830
MinimalisticVQA(IX) [46]	0.885	0.882	0.792	0.828	0.862	0.859	0.775	0.821	0.826	0.821	0.849	0.859
FAST-VQA [50]	0.880	0.880	0.781	0.813	0.859	0.854	**0.826**	0.845	0.730	0.747	0.838	0.849
DOVER [51]	0.878	0.866	0.782	0.813	0.874	0.869	0.817	0.840	0.771	0.781	0.842	0.845
Q-Align [52]	0.886	0.884	0.761	0.822	0.876	0.878	0.783	0.819	0.834	0.846	0.844	0.861
Competing Methods: LMM-based												
VQA^2 [16]	0.878	0.872	0.794	0.821	0.881	0.880	0.785	0.830	0.811	0.823	0.847	0.854
VQAThinker [2]	0.883	0.880	0.798	0.834	0.881	0.884	0.808	**0.847**	**0.860**	0.863	0.855	0.866
Our Weak-to-Strong Methods												
(I): Single teacher supervision *as baseline*	0.879	0.878	0.794	0.826	0.869	0.871	0.786	0.822	0.843	0.846	0.849	0.859
(II): (I) + Ensembling homogeneous teachers	0.883	0.877	0.804	0.829	0.883	0.876	0.799	0.830	0.843	0.845	0.856	0.860
(III): (II) + Integrating heterogeneous teachers	0.886	0.880	0.803	0.830	0.891	0.888	0.797	0.832	0.845	0.849	0.858	0.863
(IV): (III) + Confidence loss	0.885	0.881	0.803	0.831	0.890	0.891	0.797	0.833	0.849	0.856	0.857	0.865
(V): (IV) + Iterative stage W2S training	0.886	0.883	0.803	0.834	0.898	0.897	0.810	0.841	0.858	**0.864**	0.860	0.868
(VI): (V) + Iterative stage W2S training	**0.893**	**0.889**	**0.807**	**0.837**	**0.902**	**0.901**	0.818	0.846	0.852	0.858	**0.865**	**0.872**

Out of Distribution Datasets	LIVE-YT-Gaming		CGVDS		LIVE-YT-HFR		Waterloo-IVC-4K		KVQ		Overall	
# of videos	600		357		480		1,200		2,926		-	
Methods	**SRCC**	**PLCC**	**SRCC**	**PLCC**	**SRCC**	**PLCC**	**SRCC**	**PLCC**	**SRCC**	**PLCC**	**SRCC**	**PLCC**
Competing Methods: Teachers												
MinimalisticVQA(VII) [46]	0.596	0.682	0.681	0.733	0.061	0.130	0.275	0.338	0.604	0.659	0.490	0.551
MinimalisticVQA(IX) [46]	0.686	0.746	0.797	0.816	0.301	0.388	0.459	0.502	0.615	0.661	0.574	0.622
FAST-VQA [50]	0.631	0.677	0.725	0.747	0.326	0.415	0.327	0.363	0.518	0.526	0.486	0.512
DOVER [51]	0.647	0.728	0.694	0.747	0.360	0.465	0.368	0.418	0.559	0.593	0.519	0.569
Q-Align [52]	0.611	0.681	0.756	0.798	0.329	0.342	0.414	0.497	0.613	0.655	0.555	0.606
Competing Methods: LMM-based												
VQA^2 [16]	0.613	0.698	0.656	0.741	0.332	0.413	0.415	0.474	0.678	0.689	0.583	0.623
VQAThinker [2]	**0.767**	**0.806**	**0.856**	**0.845**	0.528	0.610	0.573	0.624	0.586	0.626	0.615	0.658
Our Weak-to-Strong Methods												
(I) - *as baseline*: Single teacher supervision	0.687	0.748	0.763	0.810	0.383	0.461	0.459	0.515	0.638	0.676	0.591	0.639
(II): (I) + Ensembling homogeneous teachers	0.688	0.756	0.769	0.808	0.456	0.497	0.455	0.502	0.649	0.682	0.602	0.643
(III): (II) + Integrating heterogeneous teachers	0.697	0.752	0.799	0.829	0.481	0.525	0.552	0.614	0.690	0.725	0.650	0.693
(IV): (III) + Confidence loss	0.708	0.763	0.796	0.829	0.523	0.606	0.579	0.643	0.713	0.742	0.672	0.717
(V): (IV) + Iterative stage W2S training	0.711	0.770	0.807	0.831	0.606	0.678	0.657	0.737	0.759	0.782	0.722	0.765
(VI): (V) + Iterative stage W2S training	0.723	0.776	0.799	0.828	**0.683**	**0.749**	**0.698**	**0.758**	**0.772**	**0.807**	**0.745**	**0.789**

4.4. Experimental Results

We present the experimental results in Table 2, highlighting five progressively enhanced models of our method: models (I)–(IV) incrementally add components in **Stage 1**, while model (V) and model (VI) introduce iterative training in **Stage 2** and **Stage 3**, respectively. We analyze them from the following aspects:

Ensembling Homogeneous Teachers. Compared with single-teacher supervision, we find that ensembling multiple teachers yields stronger student models that outperform all individual teachers as well as their corresponding students. This result further highlights the weak-to-strong effect in VQA and shows that improving the quality of teacher supervision amplifies this effect, consistent with prior findings.

Integrating Heterogeneous Teachers. We incorporate synthetic distortion simulators as specialized VQA models to extend the capability of the teacher ensemble. With synthetic distortion pairs, the student model achieves consistent improvements across all benchmarks, yielding marginal gains on in-domain datasets and substantial enhancements on OOD benchmarks. These results demonstrate that incorporating diverse VQA models as teachers enables joint supervision that consistently fosters more generalizable quality assessment.

Confidence Loss. Incorporating $\mathcal{L}_{\text{conf}}$ yields clear gains on OOD datasets. This indicates that confidence loss mitigates the adverse impact of noisy weak labels and enables the student to reinforce its own reliable predictions.

Iterative W2S Training. We observe consistent improvements across both in-domain and OOD datasets as the student progresses through three iterative training stages. This provides strong empirical evidence that our iterative

Table 3. Training and inference of our teacher and student models.

Model	Training Time (8×H200, hours)			Inference Time (2×RTX3090, seconds)					
	Stage 1	Stage 2	Stage 3	Ours	MinimalisticVQA(VII)	MinimalisticVQA(IX)	FAST-VQA	DOVER	Q-Align
Time	19	24	29	5.97	10.68	10.38	1.31	0.91	2.46

Table 4. Ablation study on the iterative training strategy of model (V). (V-a) denotes Stage 2 training without difficult-sample selection, where the same number of new samples are randomly chosen and their pseudo-labels refined with the Stage 1 teacher. (V-b) denotes Stage 2 training with difficult-sample selection but without refining pseudo-labels from the previous stage.

Methods	**In-domain**		**OOD**	
	SRCC	**PLCC**	**SRCC**	**PLCC**
(V-a): (V) w/o difficult-sample selection	0.857	0.866	0.669	0.714
(V-b): (V) w/o label refinement	0.858	0.862	0.657	0.702
(V)	**0.860**	**0.868**	**0.722**	**0.765**

weak-to-strong strategy enhances model capacity through progressive self-teaching. Notably, substantial gains are achieved on challenging benchmarks where existing models struggle: after three iterations, relative SRCC improvements of 30.59%, 20.55%, and 8.27% are obtained on LIVE-YT-HFR, Waterloo-IVC-4K, and KVQ, respectively. As shown in Table 4, we conduct ablation studies on Stage 2 training. Without our designed iterative training strategy, no performance improvement is observed on in-domain datasets, and clear degradation appears on OOD datasets—especially when difficult samples are selected without pseudo-label refinement. These results indicate that the performance gains originate from our iterative strategy rather than using larger training data.

Comparison with SOTAs. We compare our Stage 3 student model with state-of-the-art baselines. Our model surpasses all competitors, including the five teacher models and two recent LMM-based approaches, VQA^2 [16] and VQAThinker [2]. Notably, VQA^2 is trained on over 157k labeled samples, while VQAThinker leverages reinforcement learning with advanced LMM backbones. In contrast, our weak-to-strong learning strategy achieves state-of-the-art performance without any human-labeled data, underscoring its effectiveness and practical value.

Time Complexity. We report the runtime of both our teacher and student models, with inference averaged over 1080p videos of 240 frames, as shown in Table 3. Despite involving a full three-stage pipeline—including pseudo-label generation and student training—our method remains computationally competitive: compared with traditional subjective quality assessment, our pseudo-label generation process requires notably less time and human effort, while producing more reproducible and stable quality scores and offering better scalability. Given these advantages and the strong performance achieved by our student models, the overall computational cost of our pipeline is well justified.

5. Discussion

Developing generalized VQA models remains a fundamental challenge due to the vast diversity of real-world distortions and the strong influence of video content. Supervised learning on human-labeled data cannot feasibly cover this space, highlighting the urgent need for unsupervised and weakly supervised paradigms. In this work, we demonstrate that it is possible to learn from weak VQA models and even surpass their performance. Building on this insight, we propose a framework that integrates diverse homogeneous and heterogeneous VQA teachers through a learning-to-rank formulation, and further enhances generalization via an iterative W2S training strategy, where progressively stronger students are recycled as new teachers. This design enables cumulative transfer of knowledge beyond any single teacher and drives the model's self-evolution toward increasingly generalized quality assessment.

Looking forward, this paradigm suggests a pathway toward scalable VQA foundation models. The community can leverage a broad spectrum of supervision sources, leveraging expert-domain VQA models (*e.g.*, VMAF for video compression), utilizing powerful LMMs with carefully designed prompt engineering, and employing text-to-video generation algorithms to synthesize videos of varying quality through specified prompts, while simultaneously exploring more effective weak-label ensemble mechanisms to better unify these diverse supervisory signals. By unifying these heterogeneous signals, future research may move toward constructing foundation models for VQA that generalize across content domains, distortion types, and application scenarios—ultimately serving as universal quality assessors for both natural and generative videos.

6. Conclusion

This paper introduces a weak-to-strong (W2S) paradigm for video quality assessment that leverages multiple weak teachers and iterative self-teaching to train stronger students without relying on human annotations. Through the integration of homogeneous and heterogeneous teachers under a ranking-based formulation, and the use of iterative W2S training, our approach consistently surpasses the teacher models across ten benchmarks, with particularly strong gains on challenging out-of-distribution benchmarks. The results highlight the potential of W2S as a scalable and effective alternative to traditional annotation-dependent training pipelines.

7. Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62522116, Grant 62301316, Grant 62271312, and Grant 62132006, and in part by STCSM under Grant 22DZ2229005. We thank Professor Kede Ma for his helpful suggestions and discussions.

References

[1] Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*, 2023. 2, 3, 6

[2] Linhan Cao, Wei Sun, Weixia Zhang, Xiangyang Zhu, Jun Jia, Kaiwei Zhang, Dandan Zhu, Guangtao Zhai, and Xiongkuo Min. Vqathinker: Exploring generalizable and explainable video quality assessment via reinforcement learning. *arXiv preprint arXiv:2508.06051*, 2025. 7, 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. 1

[4] Pengfei Chen, Leida Li, Jinjian Wu, Weisheng Dong, and Guangming Shi. Contrastive self-supervised pre-training for video quality assessment. *IEEE transactions on image processing*, 31:458–471, 2021. 2

[5] Pengfei Chen, Leida Li, Jinjian Wu, Weisheng Dong, and Guangming Shi. Unsupervised curriculum domain adaptation for no-reference video quality assessment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5178–5187, 2021.

[6] Pengfei Chen, Leida Li, Haoliang Li, Jinjian Wu, Weisheng Dong, and Guangming Shi. Dynamic expert-knowledge ensemble for generalizable video quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(6):2577–2589, 2022. 2

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, G Heigold, S Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 1, 2

[8] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6202–6211, 2019. 3

[9] Qihang Ge, Wei Sun, Yu Zhang, Yunhao Li, Zhongpeng Ji, Fengyu Sun, Shangling Jui, Xiongkuo Min, and Guangtao Zhai. Lmm-vqa: Advancing video quality assessment with large multimodal models. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 3, 4

[10] Franz Götz-Hahn, Vlad Hosu, Hanhe Lin, and Dietmar Saupe. Konvid-150k: A dataset for no-reference video quality assessment of videos in-the-wild. *IEEE Access*, 9:72139–72160, 2021. 1

[11] Jianyuan Guo, Hanting Chen, Chengcheng Wang, Kai Han, Chang Xu, and Yunhe Wang. Vision superalignment: Weak-to-strong generalization for vision foundation models. *arXiv preprint arXiv:2402.03749*, 2024. 2, 3, 6

[12] David Hasler and Sabine E Suesstrunk. Measuring colorfulness in natural images. In *Human vision and electronic imaging VIII*, pages 87–95. SPIE, 2003. 3

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 1

[14] 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 *2017 Ninth international Conference on Quality of Multimedia experience*, pages 1–6, 2017. 1, 4

[15] ITU-T P.910. Subjective video quality assessment methods for multimedia applications, 2008. 1, 3

[16] Ziheng Jia, Zicheng Zhang, Jiaying Qian, Haoning Wu, Wei Sun, Chunyi Li, Xiaohong Liu, Weisi Lin, Guangtao Zhai, and Xiongkuo Min. Vqa ^{2}: Visual question answering for video quality assessment. *arXiv preprint arXiv:2411.03795*, 2024. 7, 8

[17] Ziheng Jia, Zicheng Zhang, Zeyu Zhang, Yingji Liang, Xiaorong Zhu, Chunyi Li, Jinliang Han, Haoning Wu, Bin Wang, Haoran Zhang, et al. Scaling-up perceptual video quality assessment. *arXiv preprint arXiv:2505.22543*, 2025. 1

[18] Janusz Konrad and Eric Dubois. Bayesian estimation of motion vector fields. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 14(09):910–927, 1992. 2

[19] 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. 3

[20] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 3

[21] Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *Proceedings of the 27th ACM international Conference on Multimedia*, pages 2351–2359, 2019. 2

[22] Zhuoran Li, Zhengfang Duanmu, Wentao Liu, and Zhou Wang. Avc, hevc, vp9, avs2 or av1?—a comparative study of state-of-the-art video encoders on 4k videos. In *Image Analysis and Recognition: 16th International Conference, ICIAR 2019, Waterloo, ON, Canada, August 27–29, 2019, Proceedings, Part I 16*, pages 162–173. Springer, 2019. 1, 4

[23] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Deepfl-iqa: Weak supervision for deep iqa feature learning. *arXiv preprint arXiv:2001.08113*, 2020. 2

[24] Wentao Liu, Zhengfang Duanmu, and Zhou Wang. End-to-end blind quality assessment of compressed videos using deep neural networks. In *ACM Multimedia*, pages 546–554, 2018. 2

[25] Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. Rankiqa: Learning from rankings for no-reference image quality assessment. In *Proceedings of the IEEE international conference on computer vision*, pages 1040–1049, 2017. 3

[26] Yuejiang Liu and Alexandre Alahi. Co-supervised learning: Improving weak-to-strong generalization with hierarchical mixture of experts. *arXiv preprint arXiv:2402.15505*, 2024. 3

[27] Yongxu Liu, Jinjian Wu, Leida Li, Weisheng Dong, Jinpeng Zhang, and Guangming Shi. Spatiotemporal representation learning for blind video quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(6): 3500–3513, 2021. 2

[28] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 1

[29] 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. 4

[30] 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. 6

[31] Pavan C Madhusudana, Xiangxu Yu, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C Bovik. Subjective and objective quality assessment of high frame rate videos. *IEEE Access*, 9:108069–108082, 2021. 1, 4

[32] Pavan C Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C Bovik. Conviqt: Contrastive video quality estimator. *IEEE Transactions on Image Processing*, 32:5138–5152, 2023. 2

[33] Xiongkuo Min, Huiyu Duan, Wei Sun, Yucheng Zhu, and Guangtao Zhai. Perceptual video quality assessment: A survey. *Science China Information Sciences*, 67(11):211301, 2024. 1

[34] Shankhanil Mitra and Rajiv Soundararajan. Knowledge guided semi-supervised learning for quality assessment of user generated videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4251–4260, 2024. 2

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

[36] Anish Mittal, Michele A Saad, and Alan C Bovik. A completely blind video integrity oracle. *IEEE Transactions on Image Processing*, 25(1):289–300, 2015. 2

[37] Niranjan D Narvekar and Lina J Karam. A no-reference image blur metric based on the cumulative probability of blur detection (cpbd). *IEEE Transactions on Image Processing*, 20(9):2678–2683, 2011. 3

[38] Juergen Pandel. Measuring of flickering artifacts in predictive coded video sequences. In *2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services*, pages 231–234. IEEE, 2008. 3

[39] Eli Peli. Contrast in complex images. *JOSA A*, 7(10):2032–2040, 1990. 3

[40] Piotr Romaniak, Lucjan Janowski, Mikolaj Leszczuk, and Zdzislaw Papir. Perceptual quality assessment for h. 264/avc compression. In *2012 IEEE Consumer Communications and Networking Conference*, pages 597–602. IEEE, 2012. 3

[41] Michele A Saad, Alan C Bovik, and Christophe Charrier. Blind prediction of natural video quality. *IEEE Transactions on image Processing*, 23(3):1352–1365, 2014. 2

[42] Avinab Saha, Yu-Chih Chen, Chase Davis, Bo Qiu, Xiaoming Wang, Rahul Gowda, Ioannis Katsavounidis, and Alan C Bovik. Study of subjective and objective quality assessment of mobile cloud gaming videos. *IEEE Transactions on Image Processing*, 32:3295–3310, 2023. 4

[43] Jitao Sang, Yuhang Wang, Jing Zhang, Yanxu Zhu, Chao Kong, Junhong Ye, Shuyu Wei, and Jinlin Xiao. Improving weak-to-strong generalization with scalable oversight and ensemble learning. *arXiv preprint arXiv:2402.00667*, 2024. 3

[44] Zaixi Shang, Yixu Chen, Yongjun Wu, Hai Wei, and Sriram Sethuraman. Subjective and objective video quality assessment of high dynamic range sports content. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 556–564, 2023. 1

[45] Zeina Sinno and Alan Conrad Bovik. Large-scale study of perceptual video quality. *IEEE Transactions on Image Processing*, 28(2):612–627, 2018. 1, 4

[46] 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*, 2024. 1, 3, 7

[47] Louis L Thurstone. A law of comparative judgment. In *Scaling*, pages 81–92. Routledge, 2017. 5

[48] Kristi Tsukida, Maya R Gupta, et al. How to analyze paired comparison data. *Department of Electrical Engineering University of Washington, Tech. Rep. UWEETR-2011-0004*, 1, 2011. 5

[49] Yilin Wang, Sasi Inguva, and Balu Adsumilli. Youtube ugc dataset for video compression research. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing*, pages 1–5. IEEE, 2019. 1, 4

[50] 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. Springer, 2022. 1, 3, 7

[51] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20144–20154, 2023. 3, 7

[52] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang,

Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. 1, 2, 3, 7

[53] Jinjian Wu, Yongxu Liu, Leida Li, Weisheng Dong, and Guangming Shi. No-reference video quality assessment with heterogeneous knowledge ensemble. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4174–4182, 2021. 2

[54] Zhenqiang Ying, Maniratnam Mandal, Deepti Ghadiyaram, and Alan Bovik. Patch-vq:'patching up'the video quality problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14019–14029, 2021. 1, 2, 3, 4

[55] Xiangxu Yu, Zhengzhong Tu, Zhenqiang Ying, Alan C Bovik, Neil Birkbeck, Yilin Wang, and Balu Adsumilli. Subjective quality assessment of user-generated content gaming videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 74–83, 2022. 1, 4

[56] 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. 3

[57] Yu Zhang, Xinbo Gao, Lihuo He, Wen Lu, and Ran He. Blind video quality assessment with weakly supervised learning and resampling strategy. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(8):2244–2255, 2018. 2

[58] Hanwei Zhu, Haoning Wu, Yixuan Li, Zicheng Zhang, Baoliang Chen, Lingyu Zhu, Yuming Fang, Guangtao Zhai, Weisi Lin, and Shiqi Wang. Adaptive image quality assessment via teaching large multimodal model to compare. *arXiv preprint arXiv:2405.19298*, 2024. 3, 5