RAM-VQA: Restoration Assisted Multi-Modality Video Quality Assessment

Pengfei Chen, Jiebin Yan, Rajiv Soundararajan, Giuseppe Valenzise, Li Cai, Leida Li

RAM-VQA: Restoration Assisted Multi-modality Video Quality Assessment

Pengfei Chen, Jiebin Yan, Rajiv Soundararajan, *Senior Member, IEEE*, Giuseppe Valenzise, *Senior Member, IEEE*, Li Cai, and Leida Li, *Senior Member, IEEE*

***Abstract*—Video Quality Assessment (VQA) strives to computationally emulate human perceptual judgments and has garnered significant attention given its widespread applicability. However, existing methodologies face two primary impediments: (1) limited proficiency in evaluating samples at quality extremes (*e.g.*, severely degraded or near-perfect videos), and (2) insufficient sensitivity to nuanced quality variations arising from a misalignment with human perceptual mechanisms. Although vision-language models offer promising semantic understanding, their reliance on visual encoders pre-trained for high-level tasks often compromises their sensitivity to low-level distortions. To surmount these challenges, we propose the Restoration-Assisted Multi-modality VQA (RAM-VQA) framework. Uniquely, our approach leverages video restoration as a proxy to explicitly model distortion-sensitive features. The framework operates through two synergistic stages: a prompt learning stage that constructs a quality-aware textual space using triple-level references (degraded, restored, and pristine) derived from the restoration process, and a dual-branch evaluation stage that integrates semantic cues with technical quality indicators via spatio-temporal differential analysis. Extensive experiments demonstrate that RAM-VQA achieves state-of-the-art performance across diverse benchmarks, exhibiting superior capability in handling extreme-quality content while ensuring robust generalization.**

***Index Terms*—Video quality assessment, Vision-language model, video restoration.**

I. Introduction

In the era of video-centric social media platforms, the exponential growth in video creation and sharing has underscored the importance of Video Quality Assessment (VQA) in optimizing Quality of Experience (QoE). VQA predicts the perceptual quality of videos, enabling the detection of low-quality content and guiding video enhancement and encoding systems, which ultimately improves visual fidelity while minimizing bandwidth consumption. Based on the availability of reference information, VQA methods are categorized into the following categories: Full-Reference (FR) VQA, Reduced-Reference (RR) VQA and No-Reference (NR) VQA. Given the impracticality of obtaining reference videos in user-generated content (UGC) scenarios [1]–[3], our work focuses on advancing NR-VQA methodologies.

Traditional VQA methods predominantly rely on hand-crafted features for quality prediction. With the advent of deep learning, CNN-based architectures became the main-

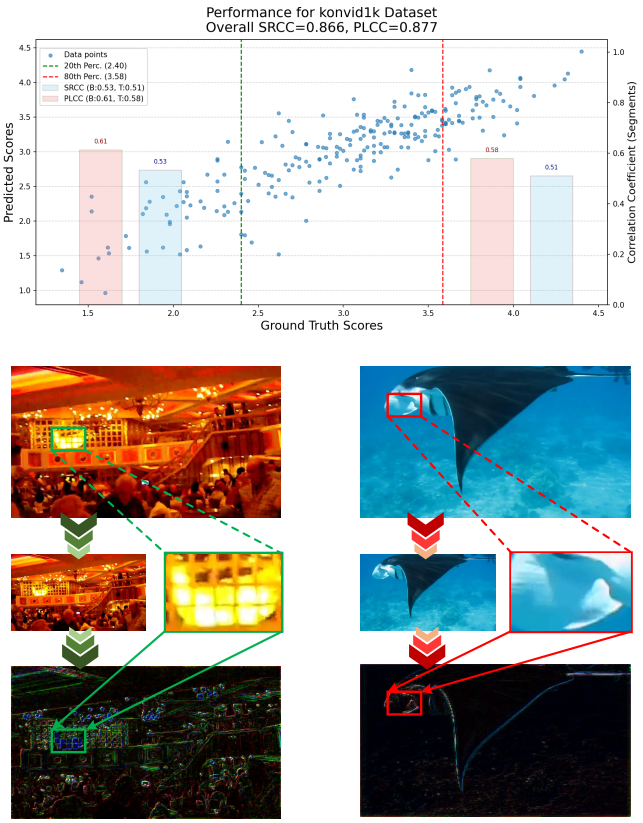


Figure 1: **Upper:** prediction performance drop of FAST-VQA [4] in extreme samples in KoNViD-1k [1]. **Lower:** residual analysis reveals information loss patterns: the left primarily captures discarded distortion artifacts (low-quality indicators), while the right predominantly contains sacrificed high-frequency details (high-quality components).

This work was supported in part by the National Natural Science Foundation of China under Grants 62301378, 62471349, 62171340 and 62461028; and in part by the China Postdoctoral Science Foundation under Grant 2024M762553; and in part by Fundamental Research Funds for the Central Universities under Grant QTZX25076 (*Corresponding author: Leida Li*).

P. Chen is with the School of Artificial Intelligence, Xidian University, Xi'an 710071, China (e-mail: chenpengfei@xidian.edu.cn).

J. Yan is with the School of Computing and Artificial Intelligence, Jiangxi University of Finance and Economics, Nanchang, Jiangxi 330032, China (e-mail: jiebinyan@foxmail.com).

R. Soundararajan is with the Department of Electrical Communication Engineering, Indian Institute of Science, Bengaluru 560012, India (e-mail: rajivs@iisc.ac.in).

G. Valenzise is with the CNRS, CentraleSupelec, Laboratoire des Signaux et Systmes, Universit Paris-Saclay, 91190 Gif-sur-Yvette, France (e-mail: giuseppe.valenzise@centralesupelec.fr).

L. Cai is with the School of Electronic and Information Engineering, Chongqing Three Gorges University, Chongqing 404100, China, China (e-mail: 20040001@sanxiau.edu.cn).

L. Li is with the School of Artificial Intelligence and State Key Laboratory of Electromechanical Integrated Manufacturing of High-Performance Electronic Equipments, Xidian University, Xi'an 710071, China (e-mail: ldli@xidian.edu.cn).

stream, estimating global video quality by applying shared convolutional filters across the entire frame. However, this global and uniform processing paradigm is inherently limited: it fails to capture the spatio-temporal variability of perceptual quality caused by regional differences in texture characteristics and distortion patterns (*e.g.*, motion blur, compression artifacts). These localized degradations may significantly deviate from the overall perceptual quality, leading to inconsistencies between predicted scores and human subjective judgments. Transformer-based methods [4] introduce attention mechanisms that can, to some extent, focus on important regions and thus partially alleviate this issue. Nonetheless, as illustrated in the upper part of Figure 1, their performance still degrades notably for videos with extremely high or low quality, even when their average accuracy over the entire quality range is satisfactory.

A major underlying cause is the intrinsic data imbalance at the boundaries of the mean opinion score (MOS) range [5], which makes it difficult for models to learn reliable representations for extreme-quality cases. To address this, a natural direction is to exploit the powerful representations of vision-language models (VLMs) [6], [7], which learn rich semantic priors from large-scale image-text corpora and have demonstrated strong generalization across tasks. In principle, such models could provide more robust and semantically informed features for VQA, especially for challenging edge cases. However, despite their ability to jointly encode visual and textual information, current VLMs still face several limitations when applied to VQA. On the visual side, the resizing operations required by standard Transformer architectures tend to suppress low-level details that are crucial for assessing visual quality. Furthermore, their visual encoders lack explicit temporal modeling, thereby ignoring essential temporal distortion clues for quality assessment. On the textual side, generic quality prompts (*e.g.*, "good"/"bad") are overly coarse and fail to capture the fine-grained, subjective nuances of human visual perception [8].

In this paper, we address the aforementioned challenges by capitalizing on the observation that resizing operations inevitably alter intrinsic perceptual properties, with particularly pronounced effects on samples at the quality extremes. As evidenced in the bottom row of Figure 1, this process simultaneously attenuates distortion patterns in low-quality content (left), while compromising fine details in high-quality material (right). Consequently, perceptual judgments become systematically biased, tending to underestimate high-quality samples while overestimating poor-quality ones. Building on this insight, we repurpose Video Super-Resolution (VSR) as a sensitive probe for quality assessment. While VSR is traditionally employed to reconstruct high-resolution frames, its training regimen renders it explicitly sensitive to low-level distortions patterns, such as blur, noise, and compression artifacts, that critically determine perceptual quality. The resulting reconstruction discrepancies, manifested as motion estimation errors or synthesized artifacts, serve as direct indicators of underlying degradation. As illustrated in Figure 1, the visual residual map between the original and VSR-reconstructed frames precisely localizes quality-sensitive regions. These regions inherently capture either complex distortion patterns in low-quality samples or missing high-frequency details in high-quality content. Since the reconstruction processes are highly consistent with how humans perceive and interpret visual stimuli in natural viewing scenarios, these differential regions provide reliable spatio-temporal guidance for quality prediction.

Based on these principles, we propose **R**estoration **A**ssisted **M**ulti-modality **V**ideo **Q**uality **A**ssessment (**RAM-VQA**) framework. The architecture leverages the Contrastive Language-Image Pre-training (CLIP) [9] model for its strong vision-language representation capacity, while innovatively incorporating restoration techniques to address its limitations in discerning technical factors. Our framework operates through two coordinated stages. In the prompt training stage, we exploit the inherent quality hierarchy among degraded, restored, and reference samples during the restoration process to optimize the vision-language alignment, enabling the model to learn text embeddings that accurately reflect human perceptual quality judgments. During quality prediction stage, input videos first undergo strategic downsampling to facilitate effective aesthetic feature learning through content understanding. The downsampled content is then reconstructed by a well-trained VSR model, with the residuals between reconstructed and original frames serving as diagnostic signals that precisely localize quality-sensitive regions and quantify degradation severity. These components further engage in temporal asymptotic interaction to simulate the temporal attention modeling inherent in human quality perception. Within our hierarchical feature learning scheme, spatio-temporal representations progressively aggregate across multiple temporal scales. The resulting technical features are fused with the aesthetic representations to form comprehensive visual embeddings, which finally align with corresponding textual representations to produce quality predictions. Our contributions are as follows:

- We propose RAM-VQA, a novel VQA approach that integrates video restoration within a multi-modal learning paradigm, which effectively guide the semantic-oriented CLIP model to develop enhanced sensitivity to critical low-level quality factors.
- We develop a quality-aware prompt learning mechanism that enables fine-grained quality discrimination by establishing optimized vision-language correspondence in CLIP's embedding space through constrained similarity optimization.
- We introduce a dual-branch feature extraction strategy, which synergistically combines semantic enhancement learning for aesthetic quality perception and temporal asymptotic interaction for technical quality analysis, significantly enhancing CLIP's ability to capture comprehensive quality-aware visual representations.
- Extensive experiments demonstrate that our method outperforms comparison approaches across multiple evaluation scenarios, with the generated quality maps showing superior consistency with human perceptual judgments.

II. Related Work

A. No-Reference Video Quality Assessment

In real-world and AIGC scenarios where pristine reference signals are unavailable [10], no-reference video quality assessment (NR-VQA) serves as the primary viable approach. Current NR-VQA methods generally fall into two categories: knowledge-driven approaches that leverage quality-related perceptual priors to extract handcrafted features, and data-driven methods that employ carefully designed networks to automatically learn quality-aware representations.

Early methods have predominantly followed knowledge-driven paradigms. For example, NIQE [11] and BRISQUE [12] pioneered the use of natural scene statistics (NSS) for image quality evaluation. Subsequent methods extended these principles to the temporal domain: VIIDEO [13] introduced natural video statistics (NVS), while V-BLIINDS [14] combined NVS with DCT-domain features. Further developments integrated complementary feature types, with TLVQM [15] incorporating both spatial high-complexity and temporal low-complexity characteristics, and VIDEVAL [16] employing feature fusion and selection to construct an ensemble prediction model.

Data-driven NR-VQA methods have achieved significant progress under the assumption of similar distribution between training and testing data [17], [18]. Representative works include VSFA [19], which incorporates temporal-memory effects in quality modeling, and RIRNet [20], which fuses spatio-temporal features across different frequency bands. Multi-task learning was explored in [21] through joint optimization of feature extraction and quality regression. Content-aware approaches include [22], which systematically analyzes content characteristics, distortion types, and compression levels, and PVQ [23], which employs patch-wise quality attributes. Recent architectural innovations include SimpleVQA [24], combining high-resolution spatial features from key frames with dense motion features, and FAST-VQA [4], introducing fragment-based sampling and enhanced Swin Transformer [25] backbones. The latter was extended in DOVER [26] with dedicated aesthetic and technical branches for comprehensive quality assessment.

B. Multimodal Models for Visual Quality Assessment

The advent of contrastive language-image pretraining has catalyzed significant advances across computer vision realms, with CLIP-based approaches demonstrating particular promise in quality assessment tasks. CLIP-IQA [8] pioneered the use of CLIP for image quality assessment (IQA), revealing its inherent capacity to perceive subjective quality attributes without task-specific fine-tuning. LIQE [27] extended this paradigm through multi-task learning, jointly optimizing vision-language encoders for quality assessment, scene classification, and distortion identification. Regarding the VQA task, methods including CLIPVQA [28], CLiF-VQA [29], and KSVQE [30] leverage CLIP-based semantic representation capabilities to model quality-related characteristics, achieving substantial performance gains in complex scenarios. MaxVQA [31] further enhanced this approach by integrating FAST-VQA features with CLIP embeddings for joint prediction of quality factors and final scores.

The emergence of large multimodal models (LMMs) has stimulated growing interest in their application to quality assessment tasks [32], [33]. In IQA Task, Q-Bench [34] explores LMM-based quality prediction by extracting softmax probabilities for quality-related tokens such as “good” and “poor”. However, LMMs without specialized fine-tuning often underperform traditional methods in tasks requiring precise visual quality understanding, primarily due to insufficient alignment between visual representations and quantitative scores. To enhance this alignment, Q-Instruct [35] fine-tunes LMMs using instructional datasets focused on visual assessment, while Q-Align [36] introduces a progressive learning syllabus that teaches quality concepts through text-defined rating levels. In VQA task, early LMM approaches suffer from temporal information loss due to frame-discrete processing. LMM-VQA [37] addresses this through spatial and temporal encoders that project visual tokens into the language space for improved modality alignment. Despite these advances, existing multimodal VQA methods remain constrained by their reliance on language-driven paradigms and labor-intensive fine-grained annotations. Our approach overcomes these limitations through quality-aware pseudo-reference generation, enabling automatic extraction of low-level technical attributes without extensive manual labeling.

III. Approach

The proposed RAM-VQA operates through two sequential stages: a prompt learning stage and a quality prediction stage, as depicted in Figure 2. In the prompt learning stage, a quality-aware prompt learning scheme is operated based on the categorized video-text pairs. Text prompts are optimized through vision-language similarity constraints across three predefined quality tiers (Figure 2(a)), enabling fine-grained quality perception. In the quality prediction stage, the input video is first downsampled to extract CLIP-based semantic representations while preserving quality-relevant features (Figure 2(b)). The downsampled version is then reconstructed using a VSR model to generate the residual information, which is combined with the original frames to perform temporal asymptotic interaction, facilitating low-level feature extraction across temporal scales (Figure 2(c)). Finally, both derived feature representations adaptively fused and aligned with the pre-learned quality prompts to generate final quality predictions (Figure 2(d)). The source code for this work is publicly available at *https://github.com/cpf0079/RAM-VQA*.

A. Prompt Learning Stage

In the CLIP architecture, the text encoder employs a multi-layer transformer architecture to encode text descriptions. To obtain effective textual prompts for VQA task, the inherent quality variations during video restoration process - spanning from severely degraded inputs to fully restored outputs - provide a natural supervisory signal for the quality assessment task. Building on this observation, we propose a novel quality-aware prompt learning framework that leverages the

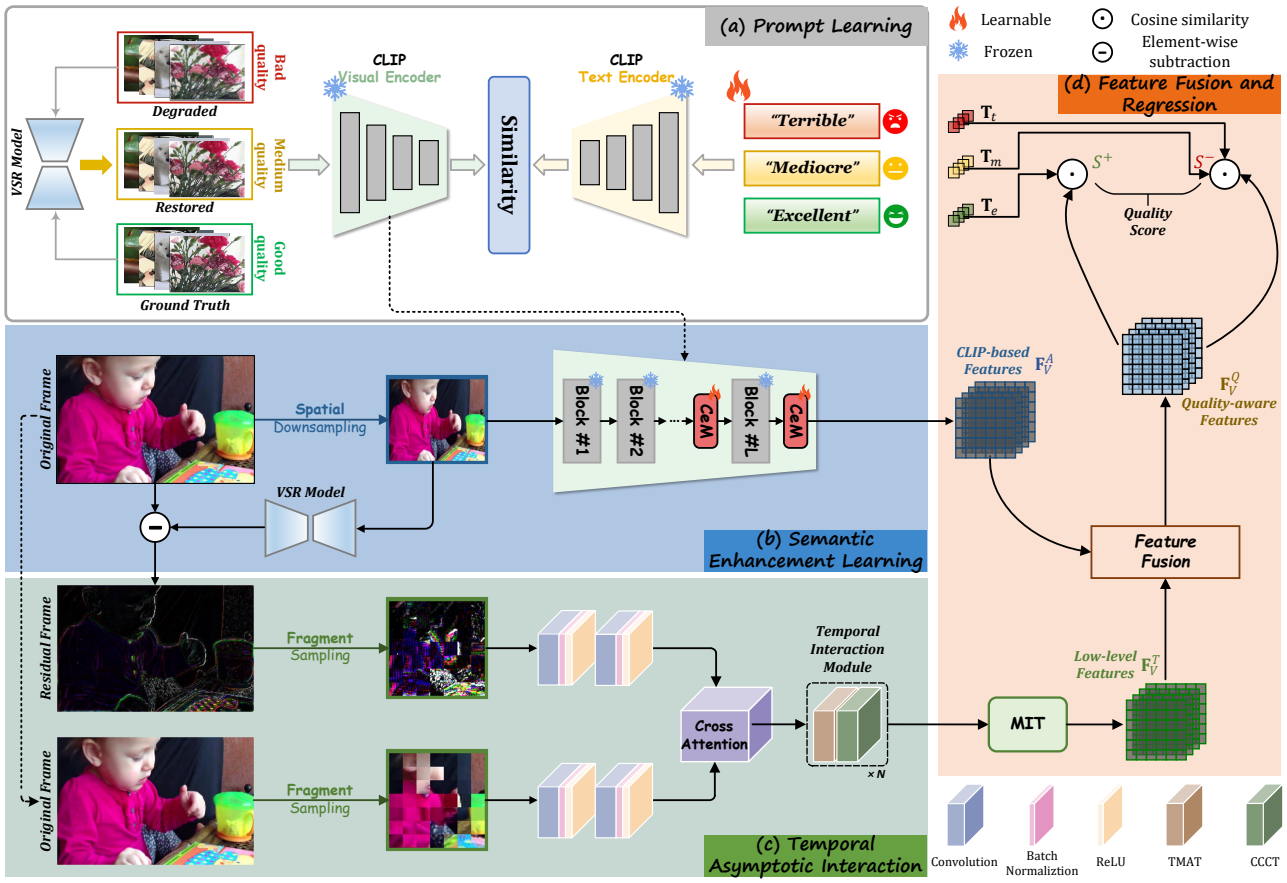


Figure 2: The overall structure of the proposed **R**estoration **A**ssisted **M**ulti-modality **V**ideo **Q**uality **A**ssessment (**RAM-VQA**), which consists of: (a) **Prompt Learning** to train textual prompts using video-text pairs categorized into three-tier of quality; (b) **Semantic Enhancement Learning** to extract aesthetic features through content understanding; (c) **Temporal Asymptotic Interaction** to extract technical features with the temporal modeling; (d) **Feature Fusion and Regression** to obtain the quality prediction through multi-modal feature alignment.

cross-modal representation capabilities of CLIP to establish discriminative text prompts for visual quality assessment.

Our approach begins by constructing a three-tier quality reference set consisting of (1) degraded videos with naive upsampling ($\mathbf{V}_b \in \mathbb{R}^{T\times H\times W\times 3}$, the input low-resolution frames were first upsampled to the same as the high-resolution ground truth using bicubic interpolation), (2) their reconstructed counterparts ($\mathbf{V}_m \in \mathbb{R}^{T\times H\times W\times 3}$) serving as medium-quality references, and (3) reference high-quality videos ($\mathbf{V}_g \in \mathbb{R}^{T\times H\times W\times 3}$). We initialize three learnable textual prompts, *i.e.*, 'terrible" ($\mathbf{T}_t \in \mathbb{R}^{N\times 512}$), "mediocre" ($\mathbf{T}_m \in \mathbb{R}^{N\times 512}$), and "excellent" ($\mathbf{T}_e \in \mathbb{R}^{N\times 512}$), corresponding to bad, medium, and good quality levels respectively. N represents the number of contextual tokens. Through an end-to-end optimization process, we jointly align these textual prompts with the visual representations obtained from CLIP visual encoder in the shared embedding space. This is achieved by minimizing a cross entropy loss $\mathcal{L}_{ce}$ that simultaneously: (i) maximizes the similarity between each prompt and its matching quality category, (ii) minimizes similarity with non-matching categories, and (iii) preserves the relative ordering of quality levels in the latent space. This learned prompt ensemble effectively captures nuanced quality distinctions while maintaining semantic consistency with human perceptual judgments. The loss can be described as:

$$\mathcal{L}_{ce} = -\frac{1}{3}\sum_{i\in\{b,m,g\}}\sum_{j\in\{e,m,t\}} y_{ij} \log\left(\frac{\exp\left(Sim\left(\mathcal{E}_{\mathcal{I}}(\mathbf{V}^{(i)}),\mathcal{E}_{\mathcal{T}}(\mathbf{T}_j)\right)\right)}{\sum_{k\in\{e,m,t\}}\exp\left(Sim\left(\mathcal{E}_{\mathcal{I}}(\mathbf{V}^{(i)}),\mathcal{E}_{\mathcal{T}}(\mathbf{T}_k)\right)\right)}\right), \tag{1}$$

where $Sim(\cdot,\cdot)$ denotes cosine similarity and y_{ij} is the classification label of the video sample $\mathbf{V}_i$, while $\mathcal{E}_{\mathcal{I}}(\cdot)$ and $\mathcal{E}_{\mathcal{T}}(\cdot)$ denote CLIP vision encoder and text encoder, respectively.

B. Quality Prediction Stage

Previous analysis of VQA reveals that perceptual quality is fundamentally governed by two complementary dimensions [26]: aesthetic factors (*e.g.*, composition, style) and technical factors (*e.g.*, distortions, artifacts). Motivated by this dichotomy, we architecturally decouple the visual feature extraction into two parallel branches, with an aesthetic branch for capturing subjective semantic attributes, and a technical branch for analyzing objective degradation patterns.

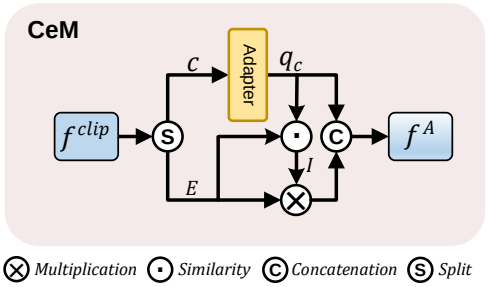


Figure 3: Illustration on how to perform content-enhanced modulation (CeM) based on the CLIP-visual encoder.

Semantic Enhancement Learning. To effectively extract aesthetic-relevant features while mitigating distortion interference, our framework first applies spatio-temporal downsampling to input videos, aiming to preserve semantic content while reducing artifact sensitivity [26]. While the CLIP model has triumphed over various downstream tasks due to its exceptional semantic awareness [7], [38], directly imposing it for aesthetic perception in VQA falls short, as the CLIP-visual encoder is not inherently quality-aware. To this end, we introduce a content-enhanced modulation (CeM) among the consecutive CLIP-visual encoder blocks. This module introduces a quality adapter to bolster the class token (*i.e.*, semantics) beyond the L-th layer, thereby achieving content enhancement that aligns with quality assessment needs.

In particular, with the output $f^{clip} = [c, E]$ of the video frame where $E \in \mathbb{R}^{N_p \times C_p}$ denotes the patch embedding, we leverage the quality adapter $\varphi(\cdot)$ to adapt the class token c into the quality-aware space. The representation sent to the next CLIP visual encoder block then can be computed with:

$$f^A = [q_c, E \cdot I], \tag{2}$$

where $q_c = \varphi(c)$, and $I \in \mathbb{R}^{N_p}$ calculates the patch-wise quality-aware semantic importance to further equip the features with the content understanding capacity, as:

$$I = \frac{q_c \cdot E^T}{\|q_c\| \|E^T\|}. \tag{3}$$

By emphasizing the importance of those patches that are most quality-related, the parts of the feature embeddings that are most related to aesthetic perception can be significantly enhanced.

Temporal Asymptotic Interaction. To compensate for the CLIP visual encoder's limited sensitivity to low-level distortion details which play a very important role in estimating the perception quality, we develop a hybrid technical feature extraction pipeline that synergistically combines: (1) pseudo-reference generation via video super-resolution $\hat{\mathbf{V}}$ and subsequent residual information computation ($\hat{\mathbf{E}} = \hat{\mathbf{V}} \ominus \mathbf{V}$) following the self-repair mechanisms of human brains for explicit distortion quantification, (2) fragmented input processing [4] for cross-attention information interaction, and (3) a novel Temporal Interaction Module (TIM) architecture featuring dual cascaded transformer networks, as shown in Figure 2(c).

Specifically, we perform "fragment" processing on the reconstructed video $\mathbf{V}^r \in \mathbb{R}^{T \times H \times W \times 3}$ as well as the original video $\mathbf{V} \in \mathbb{R}^{T \times H \times W \times 3}$, to obtain $\mathbf{V}^{r,f} \in \mathbb{R}^{T' \times H' \times W' \times 3}$ and $\mathbf{V}^f \in \mathbb{R}^{T' \times H' \times W' \times 3}$. They are fed into different convolutional layers, and the frame-level representations $\mathbf{F}^f = \{\mathbf{f}_1^f, \mathbf{f}_2^f, ..., \mathbf{f}_T^f\}$ corresponding to $\mathbf{V}^f$ is warped by $\mathbf{F}^{r,f} = \{\mathbf{f}_1^{r,f}, \mathbf{f}_2^{r,f}, ..., \mathbf{f}_T^{r,f}\}$ of $\mathbf{V}^{r,f}$ with a Multi-Head Cross Attention (MHCA), as:

$$\tilde{\mathbf{f}}_t^f = \text{MHCA}(\mathbf{f}_t^{r,f}, \mathbf{f}_t^f, \mathbf{f}_t^f). \tag{4}$$

After that, we feed the feature embeddings $\tilde{\mathbf{F}}^f = \{\tilde{\mathbf{f}}_1^f, \tilde{\mathbf{f}}_2^f, ..., \tilde{\mathbf{f}}_T^f\}$ into an L_c-block TIMs to obtain the frame-level representation h_t:

$$\mathbf{z}_t^{(l)} = \text{TIM}^{(l)}(\mathbf{z}_t^{(l-1)}), l = 1, 2, ..., L_c, \mathbf{z}_t^{(0)} = \tilde{\mathbf{f}}_t^f, \tag{5}$$

$$\mathbf{h}_t = \mathbf{z}_t^{(L_c)}[0], \tag{6}$$

where l denotes the block index in TIMs, $\mathbf{z}_t^{(L_c)}[0]$ represents the final output of the [class] token.

In order to model the short- and long-term memory effects [19], [20] during the video quality perception process and reduce the computational complexity, TIM is composed of a Temporal Mutual Attention Transformer (TMAT) for frame-level temporal modeling within a specific video clip, and a Cross-Clip Communication Transformer (CCCT) for inter-clip spatio-temporal correlation through promoting information exchange among clips.

As observed from the right part of Figure 4(a), we propose a symmetric temporal attention mechanism in TMAT that processes consecutive frame features $X \in \mathbb{R}^{2 \times N \times C}$ for local temporal modeling. The input is first decomposed into individual frames $X_1, X_2 \in \mathbb{R}^{1 \times N \times C}$, which undergo bidirectional multi-head mutual attention (MHMA) [39] to compute mutually-aligned representations. These temporally coherent features are then averaged and concatenated with the global context from multi-head self-attention (MHSA) [40], followed by progressive transformation through two layer-normalized MLP blocks with residual connections. The process can be formally expressed as:

$$X_1, X_2 = \text{Split}_0(\text{LN}(X)), \tag{7}$$

$$Y_1, Y_2 = \text{MHMA}(X_1, X_2), \text{MHMA}(X_2, X_1), \tag{8}$$

$$Y_3 = \text{MHSA}([X_1, X_2]), \tag{9}$$

$$X = \text{MLP}(\text{Concat}_2(\text{Concat}_0(Y_1, Y_2), Y_3)) + X, \tag{10}$$

$$X = \text{MLP}(\text{LN}(X)) + X, \tag{11}$$

where the subscripts in $\text{Split}(\cdot)$ and $\text{Concat}(\cdot)$ refer to the specified dimensions. To overcome the computational inefficiency $O(T^2)$ of direct pairwise frame attention, we introduce a efficient calculation method that achieves linear complexity $O(T)$. As illustrated in Figure 4(b), our approach employs a two-level hierarchical strategy: (1) at the base level, the video sequence is divided into non-overlapping 2-frame segments processed in parallel using our mutual attention formulation (Eq.(7)–(11)). (2) at the connection level, we implement a shifted window scheme [25], [41] that alternates 1-frame temporal offsets between successive layers to establish cross-segment information flow. The combination of parallel segment processing and carefully-designed temporal shifting

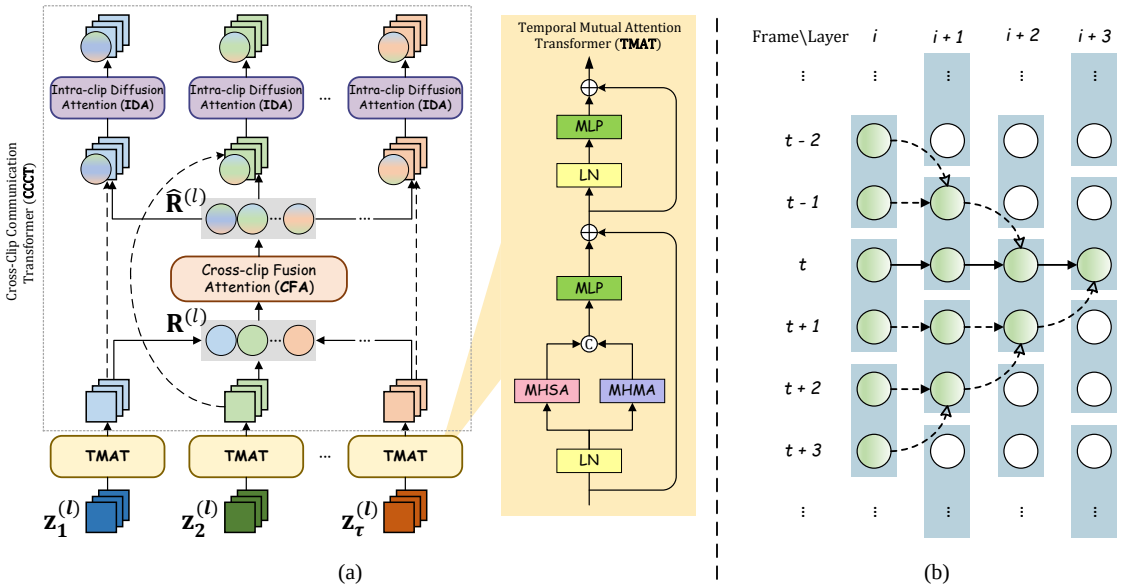


Figure 4: (a) Detailed structure of the Temporal Interaction Module (TIM). (b) Stacking of Temporal Mutual Self Attention (TMSA).

achieves an optimal balance between computational efficiency and modeling capacity, providing comprehensive temporal understanding while maintaining tractable complexity.

To facilitate inter-clip information exchange, the proposed CCCT is comprised of two complementary components: the cross-clip fusion attention (CFA) for global spatio-temporal modeling, and the intra-clip diffusion attention (IDA) for local feature refinement. They are integrated with a feed-forward network (FFN) for nonlinear transformation. As illustrated in Figure 4(a), our design employs a message token mechanism that enables dynamic information abstraction and propagation across video clips. In specific, frame-level class tokens within c-th clip at the l-th block are aggregated into compact clip-level representations through linear projection $\mathbf{r}_c^{(l)} = \beta[\mathbf{z}_{c,1}^{(l)}[0], \mathbf{z}_{c,2}^{(l)}[0], ..., \mathbf{z}_{c,\tau}^{(l)}[0]]$. These clip descriptors $\mathbf{R}^{(l)} = \{\mathbf{r}_1^{(l)}, \mathbf{r}_2^{(l)}, ..., \mathbf{r}_C^{(l)}\}$ then participate in CFA to model long-range temporal relationships across the entire video sequence. Mathematically, this process at l-th block can be expressed as:

$$\widehat{\mathbf{R}}^{(l)} = \mathbf{R}^{(l)} + \mathrm{CFA}(\mathrm{LN}(\mathbf{R}^{(l)})), \tag{12}$$

where $\widehat{\mathbf{R}}^{(l)} = \{\hat{\mathbf{r}}_1^{(l)}, \hat{\mathbf{r}}_2^{(l)}, ..., \hat{\mathbf{r}}_C^{(l)}\}$ and LN indicates layer normalization. Next, the IDA takes the clip tokens with the associated message token to learn visual representation, while the involved message token could also diffuse global spatio-temporal dependencies for learning. Mathematically, this process at l-th block can be formulated as:

$$[\hat{\mathbf{z}}_{c,t}^{(l)}, \bar{\mathbf{r}}_{c,t}^{(l)}] = [\mathbf{z}_{c,t}^{(l-1)}, \hat{\mathbf{r}}_{c,t}^{(l)}] + \mathrm{IDA}(\mathrm{LN}([\mathbf{z}_{c,t}^{(l-1)}, \hat{\mathbf{r}}_{c,t}^{(l)}])). \tag{13}$$

Finally, the feed-forward network(FFN) performs on the frame tokens as:

$$\mathbf{z}_{c,t}^{(l)} = \hat{\mathbf{z}}_{c,t}^{(l)} + \mathrm{FFN}(\mathrm{LN}(\hat{\mathbf{z}}_{c,t}^{(l)})). \tag{14}$$

Through L_c alternating layers of CFA and IDA, the CCCT progressively encodes video representations by simultaneously modeling global spatio-temporal dependencies via inter-clip information propagation, and local structural patterns through intra-clip feature refinement. This alternating attention architecture enables comprehensive video understanding while maintaining computational efficiency through the localized lifespan of message tokens.

Feature Fusion and Regression. After the temporal modeling, the Multi-frame Integration Transformer (MIT), which is constructed by the standard multi-head self-attention and feed-forward networks, takes all frame representation $\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_T\}$ as input and outputs the video-level visual representation $\mathbf{F}^T$ as following:

$$\mathbf{F}^T = \mathrm{AvgPool}(\mathrm{MIT}(\mathbf{H} + e^{temp})), \tag{15}$$

where $AvgPool$ and e^{temp} denote the average pooling and temporal position encoding, respectively. After that, the modified CLIP-based features $\mathbf{F}^A$ are fused with $\mathbf{F}^T$ through a feature fusion module, which is performed using a residual multi-layer perception (MLP):

$$\mathbf{F}^Q = MLP(\mathbf{F}^A, \mathbf{F}^T) + \mathbf{F}^A. \tag{16}$$

The final quality score is obtained by computing the semantic alignment between (1) the quality-aware text embeddings $\mathbf{T}_k$ (representing 'excellent', 'mediocre', and 'terrible' quality levels) learned in the prompt learning stage, and (2) the fused visual feature representation $\mathbf{F}^Q$ that combines both aesthetic and technical quality indicators:

$$Q = \frac{1}{T}\sum_{t=0}^{T-1} \frac{\exp(Sim(\mathbf{F}_t^Q, \mathcal{E}_{\mathcal{T}}(\mathbf{T}_e)))}{\sum_{k\in\{e,m,t\}} \exp(Sim(\mathbf{F}_t^Q, \mathcal{E}_{\mathcal{T}}(\mathbf{T}_k)))}. \tag{17}$$

IV. Experimental Analysis

A. Experimental Setups

Implementation Details. In our work, we adopt RealBasicVSR [42] as the well-trained VSR model, and ViT-B/32 as

parts of the cross-frame communication transformer. During the prompt learning stage, samples from Vimeo-90K [43] dataset are leveraged for learning discriminative text prompts.

In the quality prediction stage, to facilitate effective visual feature extraction, we implement a strategy combining spatial downsampling and temporal sparse sampling within the semantic enhancement learning. This approach is designed to preserve the high-level semantics and compositional integrity of the original videos. Specifically, we resize each video frame to a spatial resolution of 224×224 and apply uniform temporal sampling with a stride of 8 frames. For the temporal asymptotic interaction, we adopt the grid mini-patch sampling strategy introduced in [4] to obtain the fragment input. In this process, each video frame is partitioned into a 7×7 grid, and a mini-patch is randomly sampled from each grid cell. These mini-patches are then concatenated based on their original spatial coordinates. Consequently, the input is reshaped into a tensor of dimensions 32×224×224, corresponding to (7×7) fragments. It is worth noting that during the training phase, we sample a single clip consisting of 32 continuous frames, whereas for inference, we sample three such clips to ensure robust prediction.

We opt for the AdamW [44] optimizer following the cosine learning rate strategy [45] which is initialized with 1×10^{-5}. Two generic measure criteria are used for performance evaluation: Pearson linear correlation coefficient (PLCC) and spearman rank order correlation coefficient (SRCC). For these two indicators, higher values are generally equivalent to better model performance.

Datasets. We evaluate RAM-VQA on four widely-used in-the-wild VQA datasets: KoNViD-1k [1] (1,200 videos), LIVE-VQC [2] (585 videos), YouTube-UGC [3] (1,147 videos), and LSVQ [23] (3,590 videos, which are further split into $\text{LSVQ}_{\text{test}}$ and $\text{LSVQ}_{\text{1080p}}$).

B. Main Results

Pre-training Results on LSVQ. We first pre-train the proposed RAM-VQA on LSVQ and compare it with the existing advanced VQA methods. These methods have been further divided into two groups as the zero-shot methods and the supervised methods. As shown in Table I, we report the SRCC and PLCC results on both intra-dataset and generalization settings, where all experiments are conducted under 10 train-test splits. Compared with the zero-shot methods, it is desirable that RAM-VQA significantly promotes the prediction performances on all test datasets. As for the training-based comparisons, RAM-VQA showcases the advantage of attention mechanisms and maintains a significant lead in both intra-dataset and cross-dataset testing, against those without CLIP features. Compared to current best models based on CLIP structure (CLIPVQA [28], CLiF-VQA [29] and MaxVQA [31]), our method performs better through quality-aware prompt learning and semantic enhancement learning. We attained the best SRCC of 0.902 (+1.58%) and PLCC of 0.909 (+2.25%) in LSVQ_{test}. In LSVQ_{1080p}, our SRCC improved by 4.40% to reach 0.830, while our PLCC increased by 2.52% to reach 0.855. During the cross-database evaluation, RAM-VQA improved by 2.05% and 3.29% on koNViD-1k and LIVE-VQC with respect to the SRCC value, thus exemplifying its generalization ability.

Fine-tuning Results on Small Datasets. To evaluate the transfer learning capability of RAM-VQA, we fine-tune the LSVQ-pretrained model on three benchmark datasets (KoNViD-1k, LIVE-VQC, and YouTube-UGC) using 10 random 80:20 train-test splits for robust evaluation. The results in Table II demonstrate RAM-VQA's superior performance, achieving consistent improvements of +1.0% on KoNViD-1k and YouTube-UGC, and a more substantial +3.5% gain on LIVE-VQC compared to existing state-of-the-art methods. Notably, RAM-VQA outperforms other CLIP-based approaches (MaxVQA [31], CLiF-VQA [29]) due to its effective modeling of (1) semantic quality relationships through prompt learning, (2) perceptual saliency via restoration guidance, and (3) spatio-temporal dependencies using hierarchical attention. These results validate RAM-VQA's strong generalization capability across diverse datasets and its effectiveness in practical scenarios.

Evaluating on Extreme-quality Samples. Building upon our earlier discussion of saliency-guided quality assessment, we specifically evaluate RAM-VQA's effectiveness for extreme-quality samples (top and bottom 20% of the quality distribution) across all five benchmark datasets. As demonstrated in Figure 5, our method achieves consistent and substantial performance advantages over competing VQA approaches, despite incorporating no specialized modules for extreme sample processing. While conventional CLIP-based approaches (*e.g.*, CLIPVQA [28]) struggle with extreme samples due to their semantic-focused architecture, our method achieves consistent performance gains through two key mechanisms, (1) super-resolution-derived residual maps that explicitly guide attention to quality-critical regions, and (2) effective fusion of low-level perceptual cues with high-level semantic features. These results quantitatively validate our hypothesis that augmenting CLIP's representation through brain-inspired self-repair mechanisms enables more robust quality assessment across the entire spectrum, particularly for challenging extreme-quality cases where traditional methods typically exhibit significant performance degradation.

To further demonstrate the advantage of our method, Figure 6 presents comparative scatter plots evaluating RAM-VQA against FAST-VQA on KoNViD-1k's extreme-quality samples: (a) bottom 20% (lowest quality) and (b) top 20% (highest quality). From the given samples, while FAST-VQA shows characteristic performance degradation at quality extremes, RAM-VQA maintains robust accuracy throughout the quality spectrum. The introduced saliency guidance enables superior perception of quality-critical details by directing attention to perceptually salient regions. This capability proves particularly valuable for challenging real-world applications like security monitoring, where conventional methods often fail due to limited annotation quality and prevalent extreme-quality footage.

C. Ablation Studies

We conduct ablation studies to validate the utility of the core components of RAM-VQA. The experiments are conducted on

Table I: Experimental performance of the pre-trained RAM-VQA model on the LSVQ dataset on four test sets (LSVQ_{test}, LSVQ_{1080p}, KoNViD-1k and LIVE-VQC). LSVQ_{test} and LSVQ_{1080p} are used for intra-dataset testing. While KoNViD-1k and LIVE-VQC are used for cross-dataset testing.

Type/		**Intra-dataset Test Sets**				**Cross-dataset Test Sets**			
Testing Set/		LSVQ_{test}		LSVQ_{1080p}		KoNViD-1k		LIVE-VQC	
Groups	Methods	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
Zero-shot Methods	NIQE [11]	0.389	0.377	0.260	0.244	0.548	0.543	0.622	0.598
	VIIDEO [13]	0.426	0.398	0.345	0.371	0.013	0.015	0.137	0.029
	TPQI [46]	0.342	0.350	0.288	0.256	0.693	0.693	0.730	0.718
	SAQI [47]	0.598	0.585	0.534	0.522	0.760	0.760	0.794	0.784
Supervised Methods	TLVQM [15]	0.774	0.772	0.616	0.589	0.724	0.732	0.691	0.670
	VIDEVAL [16]	0.783	0.794	0.554	0.545	0.741	0.751	0.640	0.630
	RAPIQUE [48]	0.788	0.797	0.578	0.582	0.737	0.755	0.670	0.652
	VSFA [19]	0.796	0.801	0.704	0.675	0.794	0.784	0.772	0.734
	RIRNet [20]	0.816	0.813	0.722	0.708	0.781	0.789	0.780	0.748
	PVQ [23]	0.828	0.827	0.739	0.711	0.795	0.791	0.807	0.770
	BVQA [49]	0.854	0.852	0.782	0.771	0.837	0.834	0.824	0.816
	CSPT [50]	0.831	0.833	0.750	0.734	0.808	0.799	0.805	0.786
	FAST-VQA [4]	0.874	0.872	0.809	0.770	0.862	0.864	0.841	0.824
	CLIPVQA [28]	0.882	0.879	0.834	0.793	0.887	0.864	0.871	0.781
	CLiF-VQA [29]	0.887	0.886	0.832	0.790	0.874	0.877	0.855	0.834
	MaxVQA [31]	0.889	0.888	0.830	0.795	0.882	0.880	0.870	0.852
RAM-VQA (Ours)		**0.909**	**0.902**	**0.855**	**0.830**	**0.899**	**0.898**	**0.889**	**0.880**
Improvement to existing best		**2.25%**	**1.58%**	**2.52%**	**4.40%**	**1.35%**	**2.05%**	**2.07%**	**3.29%**

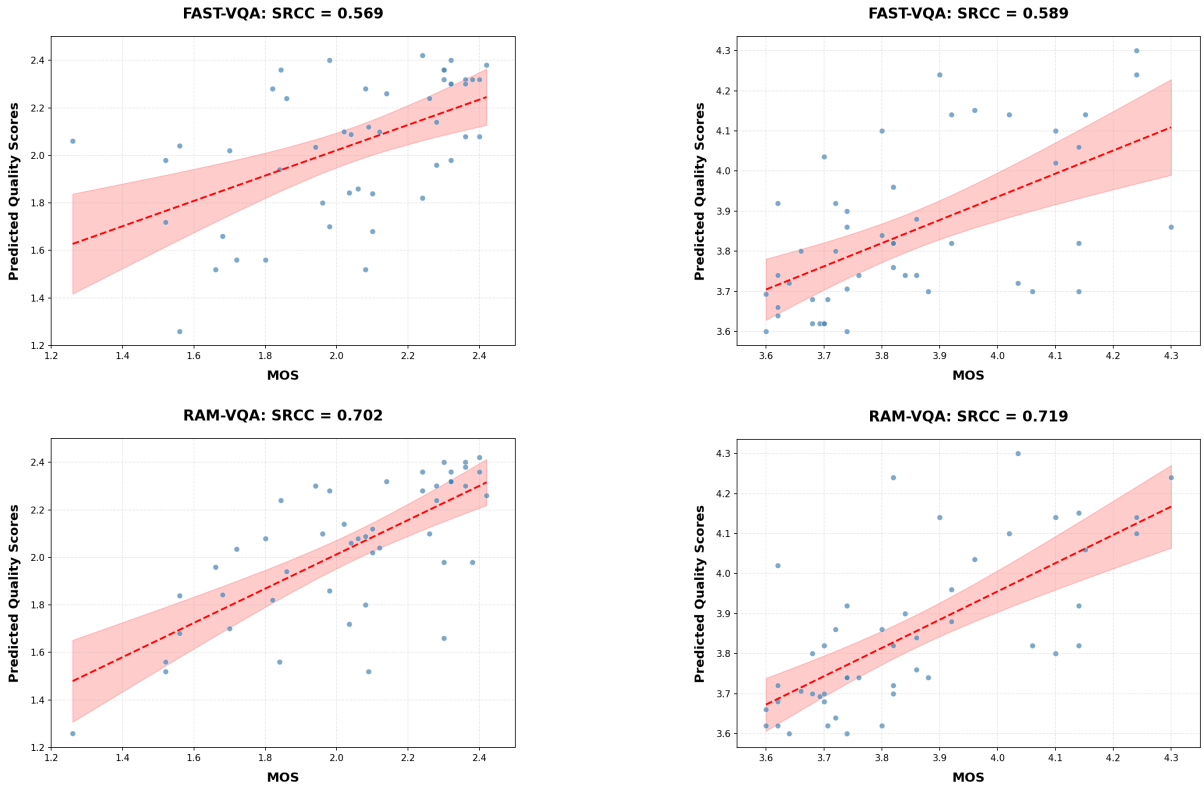
Table II: Experimental performance of the finetuned RAM-VQA model on LIVE-VQC, KoNViD and YouTube-UGC.

Database	KoNViD-1k		LIVE-VQC		YouTube-UGC		Average	
Method \ Criterion	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
TLVQM [15]	0.768	0.773	0.803	0.799	0.659	0.669	0.743	0.747
VIDEVAL [16]	0.780	0.783	0.751	0.752	0.773	0.779	0.768	0.771
VSFA [19]	0.775	0.773	0.795	0.773	0.743	0.724	0.771	0.757
RIRNet [20]	0.781	0.776	0.798	0.758	0.758	0.760	0.779	0.765
CSPT [50]	0.806	0.801	0.820	0.792	0.792	0.779	0.806	0.791
FAST-VQA [4]	0.892	0.891	0.862	0.852	0.852	0.855	0.869	0.866
CLIPVQA [28]	0.879	0.875	0.874	0.845	0.861	0.859	0.871	0.860
CLiF-VQA [29]	0.903	0.902	0.870	0.866	0.890	0.888	0.888	0.885
MaxVQA [31]	0.895	0.894	0.873	0.854	0.890	0.894	0.886	0.881
RAM-VQA (Ours)	**0.915**	**0.914**	**0.902**	**0.897**	**0.899**	**0.907**	**0.905**	**0.906**
Improvement to existing best	**1.33%**	**1.33%**	**3.20%**	**3.58%**	**1.01%**	**1.45%**	**1.91%**	**2.37%**

KoNViD-1k, LIVE-VQC and YouTube-UGC, with consistent data partitioning and parameter settings.

Effectiveness of Different Components. Our ablation study in Table III systematically evaluates the contributions of each key component in RAM-VQA, demonstrating three critical findings: (1) Quality-aware prompt learning provides substantial improvement (+0.054 SRCC on average), validating its effectiveness in capturing perceptual quality relationships; (2) Semantic enhancement learning (CeM) enhance performance by +0.033 SRCC on average; and (3) Temporal asymptotic interaction proves essential, as its removal causes significant performance degradation (-0.178 SRCC), confirming the importance of low-level temporal feature learning. Contrary to expectations, direct use of the CLIP visual encoder (denoted as the "★" setting) for feature extraction provided no measurable gain. Overall, our model achieves a significant improvement compared with the variants, which is attributed to the effectiveness of each proposed component.

Combination of Modalities. Methodologically, this section explores effective fusion of textual and visual modalities for quality prediction. We leverage the quality-graded textual features from prompt learning to construct positive and neg-



(a) (b)

Figure 6: Scatter plots for extreme samples evaluation in KoNViD-1k.

Table III: Ablation study on the effectiveness of different model components.

Strategies						Databases					
Prompt Learning	SEL	TAI				KoNViD-1k		LIVE-VQC		YouTube-UGC	
	CeM	Con.	CA	TMAT	CCT	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
✗	✓	✓	✓	✓	✓	0.865	0.863	0.855	0.847	0.861	0.846
✓	✗	✓	✓	✓	✓	0.880	0.881	0.866	0.868	0.877	0.870
✓	✓	✗	✗	✗	✗	0.770	0.749	0.742	0.738	0.709	0.697
✓	✓	★	✓	✓	✓	0.892	0.886	0.884	0.886	0.882	0.883
✓	✓	✓	✗	✓	✓	0.903	0.903	0.887	0.890	0.889	0.890
✓	✓	✓	✓	✗	✓	0.901	0.900	0.891	0.889	0.885	0.889
✓	✓	✓	✓	✓	✗	0.906	0.903	0.892	0.894	0.899	0.896
✓	✓	✓	✓	✓	✓	**0.915**	**0.914**	**0.902**	**0.897**	**0.905**	**0.906**

ative semantic pairs, evaluating their alignment with visual representations. Empirical results in Table V demonstrate that discriminative text feature selection improves model sensitivity. Based on this, we designate the top-performing textual feature as the positive quality anchor and the others as negative counterparts.

Usage Strategy of Restoration Information. We then evaluate four distinct approaches for incorporating restoration information into our quality prediction framework: (1) Direct restoration processing ("*direct*"): Utilizing only the restored video; (2) Direct concatenation ("*direct concat.*"): Concate-

Table V: Performance comparisons of RAM-VQA with diverse choices of multimodal fusion (different colors represent distinct combinations of positive and negative pairs). The performances are measured by SRCC.

Combination			KoNViD	VQC	YouTube
$\mathbf{T}_e$	$\mathbf{T}_m$	$\mathbf{T}_t$	0.914	0.897	0.906
$\mathbf{T}_e$	$\mathbf{T}_m$	$\mathbf{T}_t$	0.906	0.895	0.899
$\mathbf{T}_e$	$\mathbf{T}_m$	$\mathbf{T}_t$	0.912	0.895	0.904

Table IV: Ablation study on the choice of prompt design.

Type	Index	Prompt Configuration	PLCC↑	SRCC↑
Two levels	(a)	Totally fixed (“excellent”/“terrible”)	0.882	0.882
	(b)	Totally random	0.891	0.893
	(c)	Partial random (“excellent”/“terrible”)	0.901	0.900
Three levels	(d)	Totally fixed (“excellent”/“mediocre”/“terrible”)	0.888	0.886
	(e)	Totally random	0.897	0.896
	(f)	Partial random (“good”/“fair”/“bad”)	0.909	0.907
	(g)	Partial random (“terrible”/“mediocre”/“excellent”)	0.879	0.877
	(h)	Partial random (“excellent”/“mediocre”/“terrible”) (**Ours**)	**0.915**	**0.914**
Five levels	(i)	Totally fixed (“excellent”/“slightly better”/“mediocre”/“slightly worse”/“terrible”)	0.886	0.887
	(j)	Totally random	0.897	0.894
	(k)	Partial random (“excellent”/“good”/“fair”/“poor”/“bad”)	0.905	0.904
	(l)	Partial random (“excellent”/“slightly better”/“mediocre”/“slightly worse”/“terrible”)	0.906	0.903

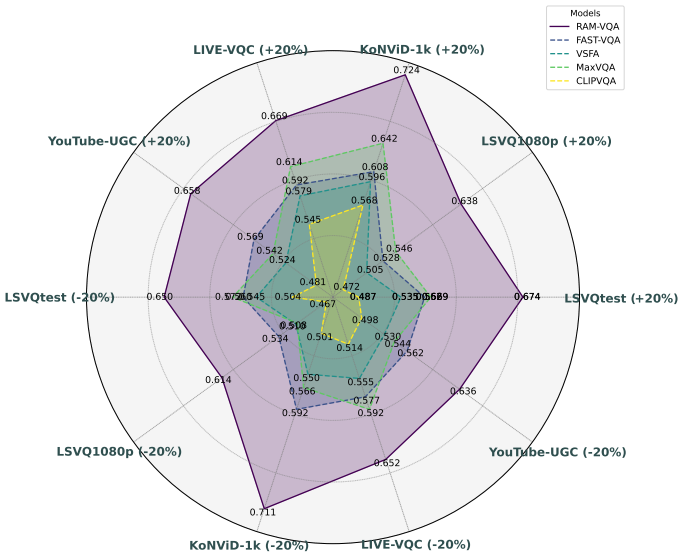


Figure 5: Performance comparison of the proposed RAM-VQA with the state-of-the-art VQA methods on extreme samples, where the evaluation metric is PLCC.

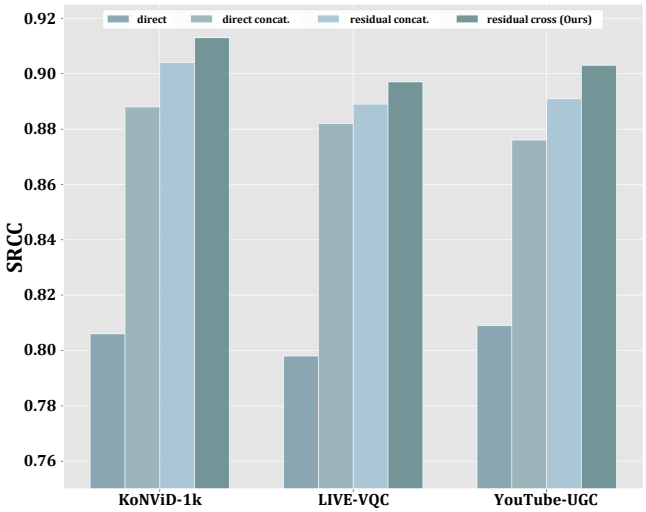


Figure 7: Performance comparison of diverse usages of restoration information, where the evaluation metric is SRCC.

nating original and restored videos; (3) Residual concatenation (“*residual concat.*”): Concatenating original video with residual maps; (4) Residual cross-attention (“*residual cross*”): Our proposed method of applying cross-attention between original and residual frames.

As demonstrated in Figure 7, the comparative analysis reveals two key findings. First, approaches combining original and restored information (settings 2-4) outperform direct restoration processing (setting 1) by significant margins, confirming that restoration data serves most effectively as complementary guidance rather than primary input. Second, The cross-attention fusion strategy (setting 4) achieves superior performance compared to concatenation methods (settings 2-3), validating our hypothesis that attention-based interaction better captures quality-relevant relationships between original content and restoration artifacts.

Choice of Prompt Design. In this study, we conduct a systematic investigation of prompt design strategies by evaluating three granularity levels (2, 3, and 5 quality tiers) and three initialization approaches: (1) *Totally fixed* that remain unchanged during training, (2) *Totally random* that are learned from scratch, and (3) *Partial random* combining fixed semantic anchors with learnable tokens (only one token is fixed as a specific word). This comprehensive analysis framework allows us to quantitatively assess how prompt granularity and initialization strategies impact video quality assessment performance.

As shown in Table IV, our comprehensive analysis of prompt design strategies reveals three key findings for effective quality assessment: (1) A three-level quality classification achieves optimal performance by balancing discrimination and granularity, outperforming both two-level (insufficient discrimination) and five-level (overly fragmented) alternatives; (2) “Partial random” initialization strategy combining fixed semantic anchors with learnable tokens demonstrates superior

effectiveness compared to "Totally fixed" or "Totally random" strategies, by better bridging CLIP's pre-trained knowledge with task-specific requirements; (3) Precise semantic formulation proves critical, as evidenced by performance degradation with vague descriptors (indexs (f, h) and indexs (k, l)) and reversed quality ordering (indexs (g, h)). These results collectively validate our optimized prompt design (index h) as achieving state-of-the-art performance through its carefully calibrated balance of granularity, initialization strategy, and semantic alignment with CLIP's embedding space.

Impact of VSR Model for Quality Perception. In this part, we explore the impact of different VSR metrics during the model design. As shown in Table VI, our systematic evaluation of VSR models for quality assessment reveals a paradoxical finding: maximal restoration fidelity does not necessarily translate to optimal quality prediction performance. As evidenced in Table VI, while state-of-the-art VSR methods like Upscale-a-Video [51] and RealViformer [52] achieve superior reconstruction quality, they underperform moderate approaches like DBVSR in our quality assessment task (SRCC: -0.33% and -0.22%). This apparent paradox arises because effective quality prediction requires preservation of diagnostically-relevant artifacts rather than their complete removal. Specifically, we identify two critical factors: (1) excessive restoration erases the subtle quality differentiators needed for accurate "mediocre" class discrimination during prompt learning, and (2) near-perfect reconstructions produce residual maps lacking discriminative power for quality perception guidance. These findings establish RealBasicVSR [42] as the optimal choice, striking the crucial balance between artifact reduction (better than MGLD-VSR [53]) and preservation of quality-discriminative features.

Table VI: Performance comparisons of RAM-VQA with different VSR usages for model training. The performances are evaluated in KoNViD-1k and measured by PLCC and SRCC.

Model	**PLCC↑**	**SRCC↑**
Upscale-a-Video (CVPR'24) [51]	0.904	0.903
MGLD-VSR (ECCV'24) [53]	0.906	0.904
RealViformer (ECCV'24) [52]	0.905	0.908
RealBasicVSR (CVPR'22) [42] (**Ours**)	**0.915**	**0.914**
DBVSR (ICCV'21) [54]	0.909	0.911

D. Computational Complexity

To evaluate the computational efficiency of the proposed RAM-VQA, we compare it with several state-of-the-art VQA methods in Table VII. Specifically, we measure the FLOPs and inference time for videos at different resolutions (540p, 720p, and 1080p). Since the semantic feature extraction and VSR process is performed offline, our model does not include these parameters during training. Due to the strategic sampling, RAM-VQA maintains consistent computational complexity across all resolutions, requiring only 862 GFLOPs and approximately 0.813 seconds per inference. This demonstrates remarkable stability compared to methods like VSFA, whose computational cost increases dramatically with resolution (from 6,440 GFLOPs at 540p to 25,712 GFLOPs at 1080p). When compared to other CLIP-based methods, RAM-VQA demonstrates superior efficiency, as it reduces FLOPs by approximately 40% compared to CLiF-VQA (1,432 GFLOPs) and improves inference speed by nearly 42% (0.813s vs 1.395s), while maintaining competitive accuracy. This balance between computational efficiency and assessment performance makes RAM-VQA particularly suitable for practical applications where both accuracy and resource constraints must be considered.

Table VII: Computational complexity comparison with other VQA methods.

Model	**540p**		**720p**		**1080p**	
	FLOPs(G)	Time(s)	FLOPs(G)	Time(s)	FLOPs(G)	Time(s)
VSFA [19]	6440	1.785	11426	2.924	25712	6.016
FAST-VQA [4]	284	0.269	284	0.270	284	0.270
DOVER [26]	283	0.337	283	0.336	283	0.337
CLIPVQA [28]	574	0.598	574	0.599	574	0.599
CLiF-VQA [29]	1432	1.395	1432	1.397	1432	1.395
RAM-VQA	862	0.812	862	0.813	862	0.813

E. Local Quality Maps

To examine the quality perception learned by RAM-VQA, we generate localized quality heatmaps following the methodology of [4]. As shown in Figure 8, our analysis of videos spanning the full quality spectrum (MOS from 1.66 to 4.16 in KoNViD-1k) reveals three key capabilities, as (1) precise texture discrimination evidenced by distinct quality predictions for sharp versus blurred regions (first and second samples) through residual fragment analysis; (2) contextual awareness demonstrated by significant quality differentiation between foreground action and background regions (third sample); and (3) effective semantic integration shown through quality predictions that align with semantically important elements (second and third samples), which demonstrates RAM-VQA is aware of and influenced by semantic information in the video, thereby demonstrating our aforementioned claims.

V. Conclusion

In this paper, we introduce RAM-VQA, a novel video quality assessment framework that capitalizes on video restoration through two complementary pathways. First, we harness the intrinsic quality hierarchy within restoration pipelines to construct discriminative quality representations via prompt learning. Second, we employ restoration residuals as explicit diagnostic signals to capture subtle distortion patterns essential for accurate quality assessment. By synergizing these restoration-derived cues, our framework achieves superior alignment with human perceptual mechanisms. Comprehensive evaluations demonstrate state-of-the-art performance across diverse benchmarks, showing particularly robust gains on content with extreme-quality variations. This work establishes video restoration as a pivotal tool for enhancing both the accuracy and interpretability of VQA systems.

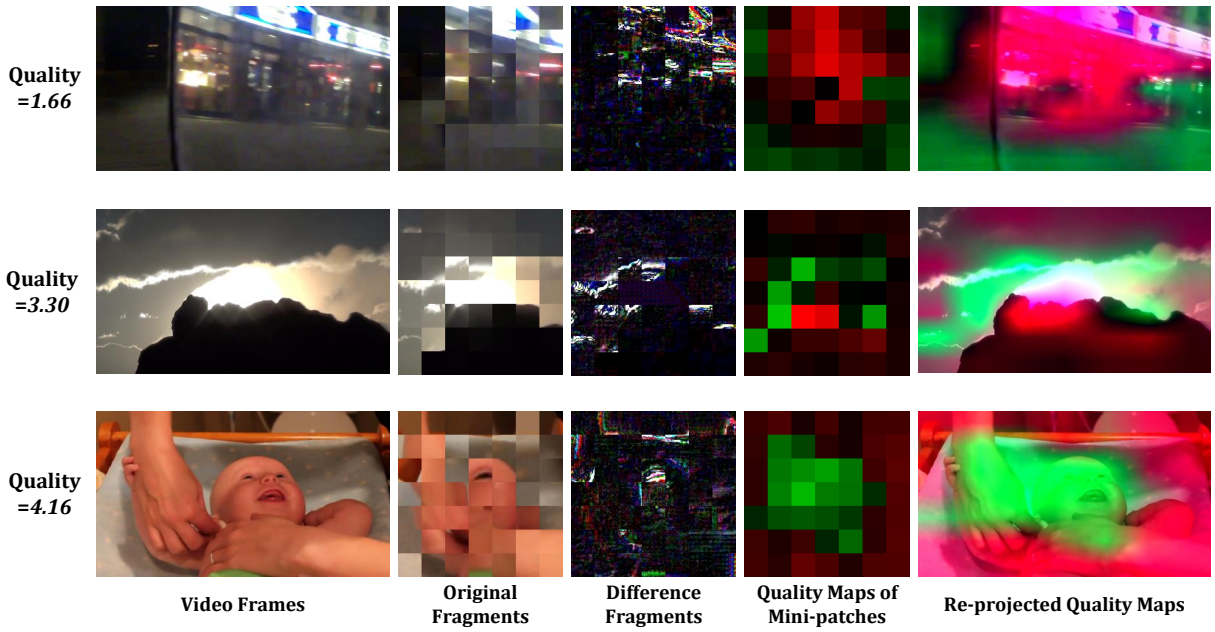


Figure 8: Video samples and their corresponding fragments, residual fragments, local quality maps and re-projected quality maps (from left to right). For quality maps, red areas refer to low predicted quality while green ones refer to high predicted quality.

REFERENCES

[1] V. Hosu, F. Hahn, M. Jenadeleh, H. Lin, H. Men, T. Szirnyi, S. Li, and D. Saupe, "The Konstanz Natural Video Database (KoNViD-1k)," in *Proc. Int. Conf. Quality Multimedia Exper. (QoMEx)*. IEEE, 2017, pp. 1–6.

[2] Z. Sinno and A. C. Bovik, "Large-Scale Study of Perceptual Video Quality," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 612–627, 2019.

[3] Y. Wang, S. Inguva, and B. Adsumilli, "YouTube UGC Dataset for Video Compression Research," in *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2019, pp. 1–5.

[4] H. Wu, C. Chen, J. Hou, L. Liao, A. Wang, W. Sun, Q. Yan, and W. Lin, "FAST-VQA: Efficient End-to-end Video Quality Assessment with Fragment Sampling," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Springer, 2022, pp. 538–554.

[5] T. Song, L. Li, D. Cheng, P. Chen, and J. Wu, "Active Learning-based Sample Selection for Label-Efficient Blind Image Quality Assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 7, pp. 5884–5896, 2023.

[6] H. Xu, G. Ghosh, P.-Y. Huang, D. Okhonko, A. Aghajanyan, F. Metze, L. Zettlemoyer, and C. Feichtenhofer, "VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 6787–6800.

[7] S. Yan, N. Dong, L. Zhang, and J. Tang, "CLIP-Driven Fine-Grained Text-Image Person Re-Identification," *IEEE Trans. Image Process.*, vol. 32, pp. 6032–6046, 2023.

[8] J. Wang, K. C. Chan, and C. C. Loy, "Exploring CLIP for Assessing the Look and Feel of Images," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 37, no. 2, 2023, pp. 2555–2563.

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning Transferable Visual Models from Natural Language Supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*. PMLR, 2021, pp. 8748–8763.

[10] P. Chen, L. Li, J. Wu, Y. Zhang, and W. Lin, "Temporal Reasoning Guided QoE Evaluation for Mobile Live Video Broadcasting," *IEEE Trans. Image Process.*, vol. 30, pp. 3279–3292, 2021.

[11] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a Completely Blind Image Quality Analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, 2012.

[12] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference Image Quality Assessment in the Spatial Domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, 2012.

[13] A. Mittal, M. A. Saad, and A. C. Bovik, "A Completely Blind Video Integrity Oracle," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 289–300, 2015.

[14] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind Prediction of Natural Video Quality," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1352–1365, 2014.

[15] J. Korhonen, "Two-Level Approach for No-Reference Consumer Video Quality Assessment," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5923–5938, 2019.

[16] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "UGC-VQA: Benchmarking Blind Video Quality Assessment for User Generated Content," *IEEE Trans. Image Process.*, vol. 30, pp. 4449–4464, 2021.

[17] P. Chen, L. Li, J. Wu, W. Dong, and G. Shi, "Unsupervised Curriculum Domain Adaptation for No-Reference Video Quality Assessment," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*. IEEE, 2021, pp. 5178–5187.

[18] P. Chen, L. Li, H. Li, J. Wu, W. Dong, and G. Shi, "Dynamic expert-knowledge ensemble for generalizable video quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 6, pp. 2577–2589, 2022.

[19] D. Li, T. Jiang, and M. Jiang, "Quality Assessment of In-the-Wild Videos," in *Proc. ACM Int. Conf. Multimedia (ACM MM)*. ACM, 2019, pp. 2351–2359.

[20] P. Chen, L. Li, L. Ma, J. Wu, and G. Shi, "RIRNet: Recurrent-In-Recurrent Network for Video Quality Assessment," in *Proc. ACM Int. Conf. Multimedia (ACM MM)*. ACM, 2020, pp. 834–842.

[21] W. Liu, Z. Duanmu, and Z. Wang, "End-to-End Blind Quality Assessment of Compressed Videos Using Deep Neural Networks," in *Proc. ACM Int. Conf. Multimedia (ACM MM)*. ACM, 2018, pp. 546–554.

[22] Y. Wang, J. Ke, H. Talebi, J. G. Yim, N. Birkbeck, B. Adsumilli, P. Milanfar, and F. Yang, "Rich Features for Perceptual Quality Assessment of UGC Videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2021, pp. 13 435–13 444.

[23] Z. Ying, M. Mandal, D. Ghadiyaram, and A. Bovik, "Patch-VQ:'Patching Up'the Video Quality Problem," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2021, pp. 14 019–14 029.

[24] W. Sun, X. Min, W. Lu, and G. Zhai, "A Deep Learning Based No-Reference Quality Assessment Model for UGC Videos," in *Proc. ACM Int. Conf. Multimedia (ACM MM)*, 2022, pp. 856–865.

[25] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video Swin Transformer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022, pp. 3202–3211.

[26] H. Wu, E. Zhang, L. Liao, C. Chen, J. Hou, A. Wang, W. Sun, Q. Yan, and W. Lin, "Exploring Video Quality Assessment on User Generated

Contents from Aesthetic and Technical Perspectives," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 20 144–20 154.
[27] W. Zhang, G. Zhai, Y. Wei, X. Yang, and K. Ma, "Blind Image Quality Assessment via Vision-Language Correspondence: A Multitask Learning Perspective," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023, pp. 14 071–14 081.
[28] F. Xing, M. Li, Y.-G. Wang, G. Zhu, and X. Cao, "CLIPVQA: Video Quality Assessment via CLIP," *IEEE Trans. Broadcast.*, vol. 71, no. 1, pp. 291–306, 2025.
[29] Y. Mi, Y. Shu, Y. Li, C. Hui, P. Zhou, and S. Liu, "CLiF-VQA: Enhancing Video Quality Assessment by Incorporating High-Level Semantic Information related to Human Feelings," in *Proc. ACM Int. Conf. Multimedia (ACM MM)*, 2024, pp. 9989–9998.
[30] Y. Lu, X. Li, Y. Pei, K. Yuan, Q. Xie, Y. Qu, M. Sun, C. Zhou, and Z. Chen, "KVQ: Kwai Video Quality Assessment for Short-Form Videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2024, pp. 25 963–25 973.
[31] H. Wu, E. Zhang, L. Liao, C. Chen, J. Hou, A. Wang, W. Sun, Q. Yan, and W. Lin, "Towards Explainable In-the-Wild Video Quality Assessment: a Database and a Language-Prompted Approach," in *Proc. ACM Int. Conf. Multimedia (ACM MM)*, 2023, pp. 1045–1054.
[32] Y. Huang, Q. Yuan, X. Sheng, Z. Yang, H. Wu, P. Chen, Y. Yang, L. Li, and W. Lin, "Aesbench: An Expert Benchmark for Multimodal Large Language Models on Image Aesthetics Perception," *arXiv preprint arXiv:2401.08276*, 2024.
[33] Y. Huang, X. Sheng, Z. Yang, Q. Yuan, Z. Duan, P. Chen, L. Li, W. Lin, and G. Shi, "Aesexpert: Towards Multi-Modality Foundation Model for Image Aesthetics Perception," in *Proc. ACM Int. Conf. Multimedia (ACM MM)*, 2024, pp. 5911–5920.
[34] H. Wu, Z. Zhang, E. Zhang, C. Chen, L. Liao, A. Wang, C. Li, W. Sun, Q. Yan, G. Zhai *et al.*, "Q-Bench: A Benchmark for General-Purpose Foundation Models on Low-level Vision," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023, pp. 10 404–10 418.
[35] H. Wu, Z. Zhang, E. Zhang, C. Chen, L. Liao, A. Wang, K. Xu, C. Li, J. Hou, G. Zhai *et al.*, "Q-Instruct: Improving Low-Level Visual Abilities for Multi-Modality Foundation Models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2024, pp. 25 490–25 500.
[36] H. Wu, Z. Zhang, W. Zhang, C. Chen, L. Liao, C. Li, Y. Gao, A. Wang, E. Zhang, W. Sun *et al.*, "Q-Align: Teaching LMMs for Visual Scoring via Discrete Text-Defined Levels," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2024, pp. 54 015–54 029.
[37] Q. Ge, W. Sun, Y. Zhang, Y. Li, Z. Ji, F. Sun, S. Jui, X. Min, and G. Zhai, "LMM-VQA: Advancing Video Quality Assessment with Large Multimodal Models," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 11, pp. 11 083–11 096, 2025.
[38] A. Sain, A. K. Bhunia, P. N. Chowdhury, S. Koley, T. Xiang, and Y.-Z. Song, "CLIP for All Things Zero-Shot Sketch-Based Image Retrieval, Fine-Grained or Not," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023, pp. 2765–2775.
[39] J. Liang, J. Cao, Y. Fan, K. Zhang, R. Ranjan, Y. Li, R. Timofte, and L. Van Gool, "Vrt: A video restoration transformer," *IEEE Trans. Image Process.*, vol. 33, pp. 2171–2182, 2024.
[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All You Need," *Proc. Advances Neural Inf. Process. Syst. (NeuraIPS)*, vol. 30, 2017.
[41] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2021, pp. 10 012–10 022.
[42] K. C. Chan, S. Zhou, X. Xu, and C. C. Loy, "Investigating Tradeoffs in Real-World Video Super-Resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022, pp. 5962–5971.
[43] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video enhancement with task-oriented flow," *Int. J. Comput. Vis.*, vol. 127, no. 8, pp. 1106–1125, 2019.
[44] I. Loshchilov, "Decoupled Weight Decay Regularization," *arXiv preprint arXiv:1711.05101*, 2017.
[45] M. U. Khattak, S. T. Wasim, M. Naseer, S. Khan, M.-H. Yang, and F. S. Khan, "Self-Regulating Prompts: Foundational Model Adaptation without Forgetting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023, pp. 15 190–15 200.
[46] L. Liao, K. Xu, H. Wu, C. Chen, W. Sun, Q. Yan, and W. Lin, "Exploring the Effectiveness of Video Perceptual Representation in Blind Video Quality Assessment," in *Proc. ACM Int. Conf. Multimedia (ACM MM)*, 2022, pp. 837–846.
[47] H. Wu, L. Liao, J. Hou, C. Chen, E. Zhang, A. Wang, W. Sun, Q. Yan, and W. Lin, "Exploring opinion-unaware video quality assessment with semantic affinity criterion," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*. IEEE, 2023, pp. 366–371.
[48] Z. Tu, X. Yu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "RAPIQUE: Rapid and Accurate Video Quality Prediction of User Generated Content," *IEEE Open J. Signal Process.*, vol. 2, pp. 425–440, 2021.
[49] B. Li, W. Zhang, M. Tian, G. Zhai, and X. Wang, "Blindly Assess Quality of In-the-wild Videos via Quality-Aware Pre-Training and Motion Perception," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 9, pp. 5944–5958, 2022.
[50] P. Chen, L. Li, J. Wu, W. Dong, and G. Shi, "Contrastive Self-Supervised Pre-Training for Video Quality Assessment," *IEEE Trans. Image Process.*, vol. 31, pp. 458–471, 2022.
[51] S. Zhou, P. Yang, J. Wang, Y. Luo, and C. C. Loy, "Upscale-a-Video: Temporal-Consistent Diffusion Model for Real-World Video Super-Resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2024, pp. 2535–2545.
[52] Y. Zhang and A. Yao, "RealViformer: Investigating Attention for Real-World Video Super-REesolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Springer, 2024, pp. 412–428.
[53] X. Yang, C. He, J. Ma, and L. Zhang, "Motion-Guided Latent Diffusion for Temporally Consistent Real-World Video Super-Resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Springer, 2024, pp. 224–242.
[54] J. Pan, H. Bai, J. Dong, J. Zhang, and J. Tang, "Deep Blind Video Super-Resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2021, pp. 4811–4820.

Pengfei Chen received the B.S. degree from Xidian University, Xian, China, in 2014, and the Ph.D. degree from China University of Mining and Technology, Xuzhou, China, in 2022. He is currently a lecturer with the School of Artificial Intelligence, Xidian University. His research interests include image/video quality assessment, video quality of experience and domain adaptation/generalization.

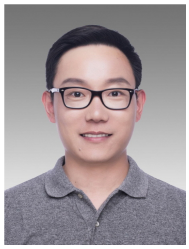
Jiebin Yan received the Ph.D. degree from Jiangxi University of Finance and Economics, Nanchang, China. He was a computer vision engineer with MT-lab, Meitu. Inc, and a research intern with MOKU Laboratory, Alibaba Group. From 2021 to 2022, he was a visiting Ph.D. student with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. From October 2024 to August 2025, he was a postdoctoral at the City University of Hong Kong, Hong Kong. He is currently an Associate Professor with the School of Computing and Artificial Intelligence, Jiangxi University of Finance and Economics, Nanchang, China. His research interests include visual quality assessment and computer vision.

Rajiv Soundararajan (Senior Member, IEEE) received the B.E. degree in Electrical and Electronics Engineering from Birla Institute of Technology and Science (BITS), Pilani, India in 2006. He received the M.S. and Ph.D. degrees in Electrical and Computer Engineering from The University of Texas at Austin, USA in 2008 and 2012 respectively. Between 2012 and 2015, he was with Qualcomm Research India, Bangalore. He is currently an Associate Professor at the Indian Institute of Science, Bangalore. He received the 2016 IEEE Circuits and Systems for Video Technology Best Paper Award and 2017 IEEE Signal Processing Letters Best Paper Award. He also received a Technology and Engineering Emmy® Award from the National Academy of Television Arts & Sciences in 2021 for the "Development of Perceptual Metrics for Video Encoding Optimization". His research interests are broadly in image and video signal processing, computer vision, machine learning and information theory.

Giuseppe Valenzise (Senior Member, IEEE) is a CNRS researcher at Universit Paris-Saclay, CNRS, CentraleSuplec, France, in the Laboratoire des Signaux et Systmes, where he leads Multimedia and Networking team. He is the Editor in Chief of the EURASIP Journal on Image and Video Processing. Giuseppe obtained his Ph.D. degree from Politecnico di Milano, Italy. In 2012, he joined the French Centre National de la Recherche Scientifique (CNRS) as a permanent researcher. His research interests span different fields of image and video processing, including traditional and learning-based image and video compression, light fields and point cloud coding, image/video quality assessment, high dynamic range imaging and applications of machine learning to image and video analysis. He has co-authored over 100 research publications in these areas. He received the EURASIP Early Career Award in 2018 for "significant contributions to video coding and analysis".

Giuseppe serves/has served as Associate Editor for IEEE Transactions on Circuits and Systems for Video Technology, IEEE Transactions on Image Processing (outstanding editorial board member award in 2022 and 2023), Elsevier Signal Processing: Image communication. He is the Chair of the Multimedia Signal Processing (MMSP) technical committee of the IEEE Signal Processing Society. Giuseppe is one of the general co-chairs of the IEEE Int. Conference on Multimedia & Expo (ICME) 2025, held in Nantes, France.

Li Cai received his BS degree from China University of Mining and Technology in 2004 and his master's degree from Chongqing University in 2010. From 2014 to 2015, he was a visiting scholar at the School of Electrical Engineering, Southeast University. He is currently a professor and the head of the Department of Electrical Engineering at Chongqing Three Gorges University. His research interests include key technologies of electric vehicle battery packs and intelligent driving technologies. He currently serves as a young editorial board member of the Battery Journal and the director of the Wanzhou District Technology Innovation Center.



Leida Li (Senior Member, IEEE) received the B.S. and Ph.D. degrees from Xidian University, Xian, China, in 2004 and 2009, respectively. From 2014 to 2015, he was a Visiting Research Fellow with the Rapid-Rich Object Search (ROSE) Lab, Nanyang Technological University, Singapore, where he was a Senior Research Fellow from 2016 to 2017. He is currently a Full Professor with the School of Artificial Intelligence, Xidian University, China. His research interests include multimedia quality assessment, computational aesthetics and visual sentiment analysis. He served as Area Chair for ACM Multimedia 2025, SPC for IJCAI 2019-2021 and 2025. He is currently an Associate Editor of IEEE Transactions on Image Processing, Journal of Visual Communication and Image Representation and EURASIP Journal on Image and Video Processing.