Temporal Inconsistency Guidance for Super-resolution Video Quality Assessment

Yixiao Li[1,5], Xiaoyuan Yang[1*], Weide Liu[2], Xin Jin[3], Xu Jia[4], Yukun Lai[5], Paul L. Rosin[5], Hantao Liu[5], Wei Zhou[5*]

[1]School of Mathematical Sciences, Beihang University, China
[2]School of Computing and Artificial Intelligence, Jiangxi University of Finance and Economics, China
[3]College of Information Science and Technology, Eastern Institute of Technology, China
[4]School of Artificial Intelligence, Dalian University of Technology, China
[5]School of Computer Science and Informatics, Cardiff University, United Kingdom
by2209112@buaa.edu.cn, xiaoyuanyang@vip.163.com, zhouw26@cardiff.ac.uk

Abstract

As super-resolution (SR) techniques introduce unique distortions that fundamentally differ from those caused by traditional degradation processes (e.g., compression), there is an increasing demand for specialized video quality assessment (VQA) methods tailored to SR-generated content. One critical factor affecting perceived quality is temporal inconsistency, which refers to irregularities between consecutive frames. However, existing VQA approaches rarely quantify this phenomenon or explicitly investigate its relationship with human perception. Moreover, SR videos exhibit amplified inconsistency levels as a result of enhancement processes. In this paper, we propose *Temporal Inconsistency Guidance for Super-resolution Video Quality Assessment (TIG-SVQA)* that underscores the critical role of temporal inconsistency in guiding the quality assessment of SR videos. We first design a perception-oriented approach to quantify frame-wise temporal inconsistency. Based on this, we introduce the Inconsistency Highlighted Spatial Module, which localizes inconsistent regions at both coarse and fine scales. Inspired by the human visual system, we further develop an Inconsistency Guided Temporal Module that performs progressive temporal feature aggregation: (1) a consistency-aware fusion stage in which a visual memory capacity block adaptively determines the information load of each temporal segment based on inconsistency levels, and (2) an informative filtering stage for emphasizing quality-related features. Extensive experiments on both single-frame and multi-frame SR video scenarios demonstrate that our method significantly outperforms state-of-the-art VQA approaches.

Code —
https://github.com/Lighting-YXLI/TIG-SVQA-main

Introduction

The rapid advancement of video processing and transmission technologies has led to an explosion of diverse video content, profoundly influencing daily life. Numerous VQA datasets have been developed, each focusing on different types of distortions such as compression distortions, user-generated degradations, and super-resolution (SR) artifacts (Hosu et al. 2017; Sinno and Bovik 2019; Xu et al. 2021; Zhou et al. 2024). Among them, SR introduces a unique class of artifacts, including hallucinated textures and temporal flickering, which differ substantially from traditional distortions. Although several SR-oriented VQA datasets have recently emerged, the development of SR-specific VQA models remains a pressing challenge due to the distinctive characteristics of SR-generated content.

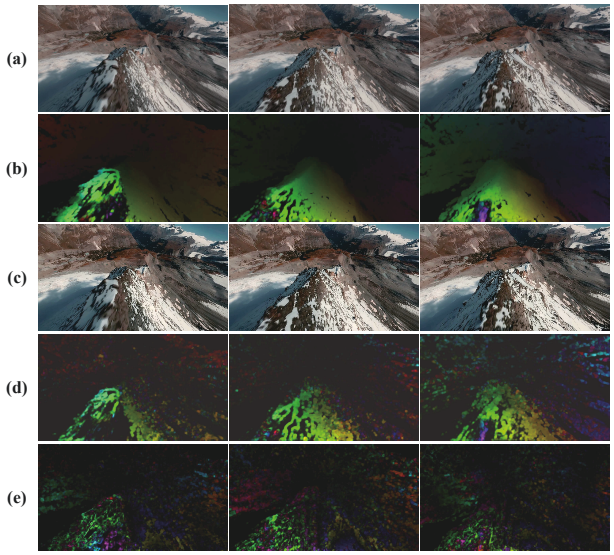

Figure 1: Visualizations for comparing temporal inconsistency with motion. Rows (a) and (c) show consecutive frames of a reference video and SR video, respectively. (b) and (d) are the optical flow of (a) and (c), respectively. (e) is the **temporal inconsistency** information for (c).

Due to the increasing demand for evaluating the quality of various video distortions, a wide range of VQA methods (Ebenezer et al. 2020; Chen et al. 2022; You and Lin 2022; Sun et al. 2022; Bi et al. 2024; Zhu et al. 2023; Korhonen 2019; Tu et al. 2021; Banitalebi-Dehkordi et al. 2020) have gained significant attention. Among the critical factors

*The corresponding authors.

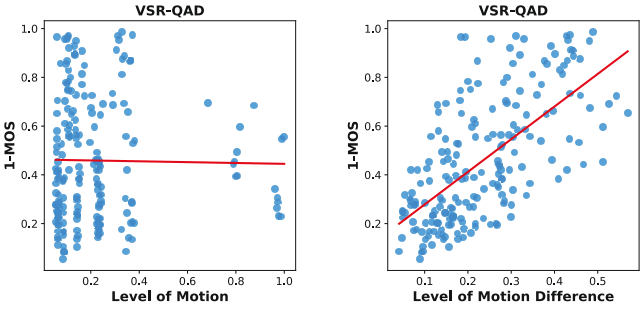

Figure 2: Correlation and performance comparison between motion and motion difference for model guidance. We analyze the correlation (SRCC) between perceptual quality (1–MOS) and motion or motion difference based on video complexity (Eq. 7 and Eq. 8). Performance comparison shows that motion achieves SRCC/PLCC of 0.885/0.913, while motion difference reaches 0.939/0.942 on the Combined-VSR dataset.

influencing video quality, temporal inconsistency refers to irregularities or disruptions in dynamic scenes (such as motion artifacts, abrupt transitions, or unnatural visual changes) that deviate from the expected smooth flow of visual content across consecutive frames. It plays a pivotal role. Recent VQA methods typically model temporal relationships using techniques such as frame differencing (Zhu et al. 2023; Bampis et al. 2017), optical flow analysis (Zhu et al. 2023), temporal slicing (Zhou et al. 2024; Ebenezer et al. 2020), natural scene statistics (Saad, Bovik, and Charrier 2014; Li, Jiang, and Jiang 2019), and 3D convolutional neural networks (Li et al. 2022; Wen et al. 2024) applied directly to distorted videos. However, none of these approaches explicitly quantify temporal inconsistency levels or examine their correlation with human perception. Moreover, the rapid advancement of SR technologies (Imani, Islam, and Wong 2022; Chan et al. 2022) has introduced additional challenges, as enhancement processes often amplify temporal inconsistencies, making it even more critical to address this issue in SR-specific VQA.

Given the potential to explore temporal inconsistency in SR-specific VQA, we begin by quantifying this phenomenon and investigating its correlation with perceptual quality. Temporal inconsistency in SR videos often arises during motion transitions (Zhang, Liu, and Xiong 2020), prompting us to analyze the relationship between motion complexity and the mean opinion score (MOS). As shown in the left part of Figure 2, motion complexity exhibits a weak correlation with perceptual quality, likely due to its strong dependence on the intrinsic complexity of scene content. This suggests that scene content can mask temporal artifacts, making motion-based signals less reliable. To mitigate this masking effect, we compute the difference in motion information between the distorted video and its reference, which we refer to as the temporal inconsistency information. As illustrated in the right part of Figure 2, the temporal inconsistency correlates strongly with perceptual quality, reinforcing the value of temporal inconsistency as a perceptual cue. Visualizations of this temporal inconsistency information are further provided in Figure 1.

Therefore, incorporating temporal inconsistency information as a guiding signal in SR-specific VQA may be beneficial for improving alignment with human perceptual preferences. To this end, we propose the **Temporal Inconsistency Guidance for Super-resolution Video Quality Assessment (TIG-SVQA)** framework, which integrates temporal inconsistency information to guide both spatial and temporal feature modeling.

To process spatial information, we highlight frame-by-frame inconsistent regions using the computed temporal inconsistency map. Spatial features are extracted at two granularities. At the coarse level, we introduce a Deformable Window Super-Attention (DW-SA) Transformer to capture large-scale inconsistencies caused by major scene transitions or fast motion, leveraging the global modeling capability of Transformers (Liu et al. 2021). At the fine level, we adopt CNNs to detect subtle distortions arising from minor motion or slow transitions, utilizing their strength in capturing local details (He et al. 2016). The two levels of features are subsequently fused to form the final spatial representation.

For temporal modeling, we design a two-stage aggregation including Consistency-aware Fusion and Informative Filtering. The first stage is inspired by the visual working memory (VWM) mechanism in human perception. While previous works (Banitalebi-Dehkordi et al. 2020; Li, Jiang, and Jiang 2019) have considered VWM, they often ignore its capacity limitation, a key property supported by cognitive studies (Cowan 2001; Zhang and Luck 2008; Schmidt et al. 2004; Dempere-Marco, Melcher, and Deco 2012). To address this, we propose a visual memory capacity block that dynamically allocates memory resources across time segments based on the inconsistency intensity. In the second stage, we perform temporal informative feature selection to retain quality-related features, ultimately enabling perceptually aligned, cross-time-scale quality prediction.

The main contributions of this paper are summarized as follows:

1. We propose the Temporal Inconsistency Guidance for Super-resolution Video Quality Assessment (TIG-SVQA) method, developing temporal inconsistency guidance for quality prediction and validating both its rationale and effectiveness.

2. We propose the Inconsistency Highlighted Spatial Module (IHSM), designed to decouple temporal inconsistency and highlight pixel-level regions exhibiting temporal irregularities across two spatial granularities.

3. We propose the Inconsistency Guided Temporal Module (IGTM), which includes the consistency-aware fusion stage and the informative filtering stage. In particular, a visual memory capacity block is proposed to dynamically allocate memory thresholds for temporal feature segmentation based on detected inconsistency levels.

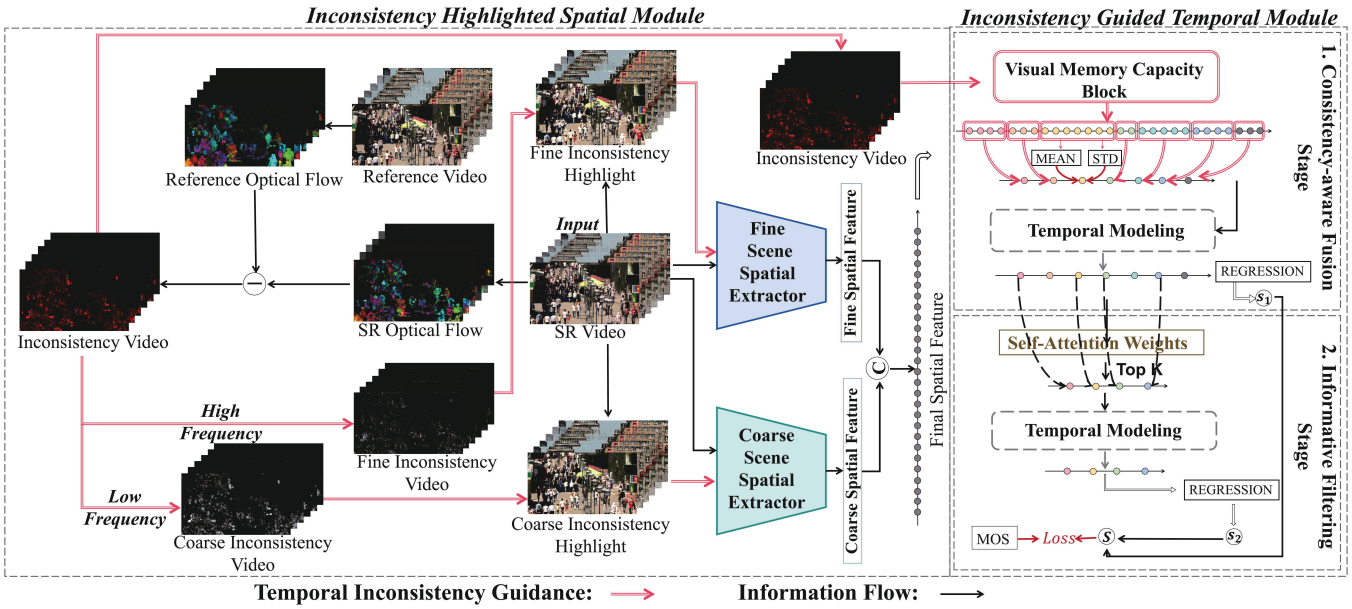

Figure 3: The framework of the proposed TIG-SVQA. The spatial module computes temporal inconsistency and applies pixel-level weighting at both coarse and fine granularities to emphasize inconsistent regions. The temporal module involves two stages: Consistency-aware Fusion and Informative Filtering.

Related Work

Temporal Relationship Modeling in VQA

Contemporary VQA methods explore temporal modeling through five main strategies: frame difference (Zhu et al. 2023; Bampis et al. 2017), optical flow (Zhu et al. 2023), spatio-temporal slicing (Zhou et al. 2024; Ebenezer et al. 2020), 3D-CNNs (Li et al. 2022; Wen et al. 2024), and multi-level features (Saad, Bovik, and Charrier 2014; Korhonen 2019). Traditional approaches often rely on handcrafted features, for example, BLIINDS (Saad, Bovik, and Charrier 2014) assesses motion coherence and global motion, while TLVQM (Korhonen 2019) uses frame complexity and temporal feature statistics. Recent learning-based methods such as STI-VQA (Zhu et al. 2023), MBVQA (Wen et al. 2024), and FAST-VQA (Wu et al. 2022) improve temporal modeling but are not optimized for SR scenarios, where upsampling can amplify temporal inconsistencies. For SR-specific VQA, only a few studies such as ERQA (Kirillova et al. 2021), Zhou et al.'s SR-VSR (Zhou et al. 2024), and SR-VQA (Cao et al. 2024) propose CNN- and saliency-based approaches, yet temporal inconsistency guidance remains largely unexplored.

Visual Working Memory Mechanism in VQA

Recent research in psychology has highlighted the crucial role of visual working memory (VWM) in shaping temporal perception (Sigala, Kaldy, and Reynolds 2022; Postle 2015). Although there are existing VQA methods considering memory modeling (Santangelo and Macaluso 2013; Banitalebi-Dehkordi et al. 2020), the application of VWM in VQA remains relatively underexplored. A defining characteristic of VWM is its limited storage capacity, which constrains both the retention and processing of visual information (Sigala, Kaldy, and Reynolds 2022; Zhang and Luck 2008). Typically, this capacity is restricted to approximately 3 to 7 visual objects (Cowan 2001; Zhang and Luck 2008), and memory accuracy tends to decline as object complexity increases (Schmidt et al. 2004; Dempere-Marco, Melcher, and Deco 2012). Further research has shown that VWM dynamically allocates its limited resources by prioritizing salient or critical scenes, which directly impacts how temporal relation is processed and modeled (Dube and Al-Aidroos 2019; Hajonides et al. 2020).

Proposed Method

In this work, we propose a **Temporal Inconsistency Guidance for Super-resolution Video Quality Assessment** method. As illustrated in Figure 3, the temporal inconsistency information guides the model learning in both spatial and temporal processing.

Inconsistency Highlighted Spatial Module (IHSM)

In this section, we introduce the inconsistency guidance for the spatial dimension. We first calculate the temporally inconsistent areas of SR videos. Taking a pair of SR video and the corresponding reference $\mathrm{V_D}$, $\mathrm{V_R} \in \mathbb{R}^{\mathrm{F}\times\mathrm{W}\times\mathrm{H}\times 3}$ as inputs, where F is the total number of frames, and $\mathrm{W}, \mathrm{H}, 3$ denote the width, height, and number of channels of each frame, respectively. The temporal inconsistency information $\mathrm{V_I}$ is captured as follows:

$$\mathrm{V_I} = \|(\mathrm{OF}(\mathrm{V_R}) - \mathrm{OF}(\mathrm{V_D}))\|_2, \mathrm{V_D} \in \text{SR videos}, \quad (1)$$

where $\mathrm{OF}(\cdot)$ refers to optical flow computation. $\|(\cdot)\|_2$ is 2-norm.

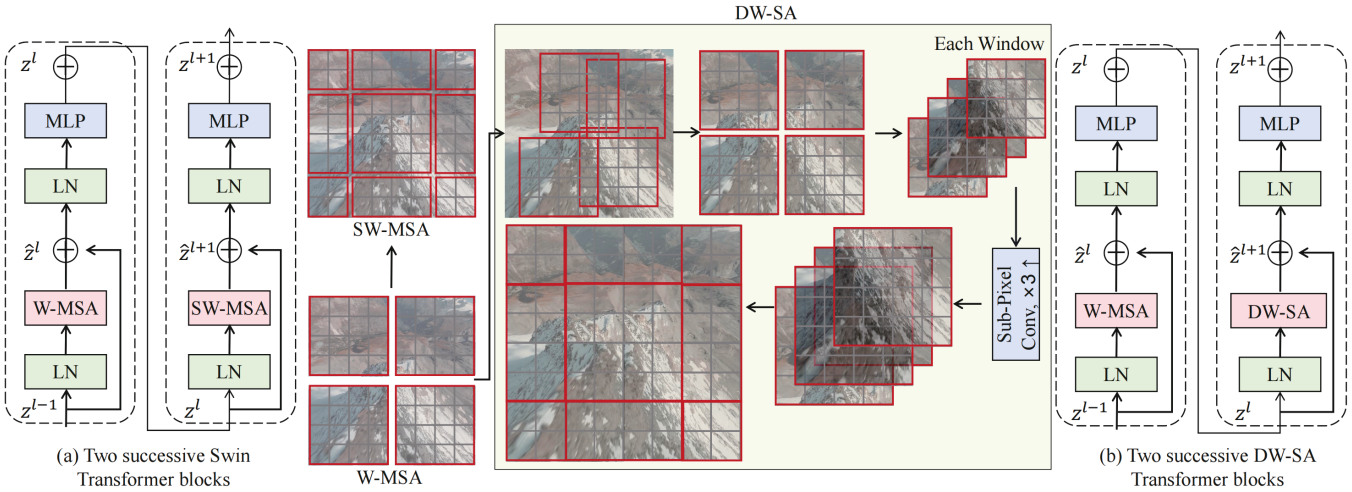

Figure 4: The details of the proposed Deformable Window Super-Attention (DW-SA) Transformer block, which adaptively adjusts window locations, up-samples features within each window, and then shifts the windows.

To capture temporal inconsistencies at different scales, we decouple the inconsistency information into coarse and fine granularities. The **coarse-grain** focuses on fast-changing regions such as scene cuts or rapid motions, obtained by applying a Gaussian low-pass filter to the video $\mathrm{V_I}$ in the frequency domain:

$$\mathrm{V_I^C} = \mathcal{F}^{-1}(\mathrm{H_L} \cdot \mathcal{F}(\mathrm{V_I})), \tag{2}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the Fourier and inverse Fourier transforms, and $\mathrm{H_L}$ is a Gaussian low-pass filter with a cutoff frequency set to 5% of the frame's longer dimension (Gonzalez, Woods, and Masters 2009).

The **fine-grain** captures subtle inconsistencies from slow transitions or minor motions by applying a Gaussian high-pass filter:

$$\mathrm{V_I^F} = \mathcal{F}^{-1}(\mathrm{H_H} \cdot \mathcal{F}(\mathrm{V_I})), \quad \mathrm{H_H} = 1 - \mathrm{H_L}. \tag{3}$$

Both filtered outputs are normalized and used to highlight inconsistency regions in the SR video V_D:

$$\{\hat{\mathrm{V}}_\mathrm{D}^{\ \mathrm{C}}, \hat{\mathrm{V}}_\mathrm{D}^{\ \mathrm{F}}\} = \mathrm{Norm}(\{\mathrm{V_I^C}, \mathrm{V_I^F}\}) \times \mathrm{V_D} + \mathrm{V_D}, \tag{4}$$

where $\mathrm{Norm}(\cdot)$ scales values to $[0, 1]$, and the outputs $\hat{\mathrm{V}}_\mathrm{D}^{\ \mathrm{C}}$ and $\hat{\mathrm{V}}_\mathrm{D}^{\ \mathrm{F}}$ are normalized to the $[0, 255]$ range.

For the coarse scene spatial extractor $\mathrm{Extractor_C}$, we propose a Transformer-based model, which has proven highly effective in capturing long-range dependencies (Liu et al. 2021). Specifically, we propose the Deformable Window Super-Attention (DW-SA) Transformer block to enhance the window adaptation. Considering that deformable technology is empirically designed in the later stages of the network (Dai et al. 2017; Zhu et al. 2019), the proposed DW-SA-T block replaces the Swin-T blocks in the third stage rather than the early stages. As shown in Figure 4, the DW-SA-T block introduces learnable offsets for each window. Since sub-pixel convolution was designed for SR image representation, each window is upsampled using sub-pixel convolution (Shi et al. 2016) and then shifted following the protocol of Swin-T. The consecutive DW-SA Transformer blocks are computed as:

$$\begin{aligned}
\hat{\mathbf{z}}^l &= \text{W-MSA}\left(\mathrm{LN}\left(\mathbf{z}^{l-1}\right)\right) + \mathbf{z}^{l-1} \\
\mathbf{z}^l &= \mathrm{MLP}\left(\mathrm{LN}\left(\hat{\mathbf{z}}^l\right)\right) + \hat{\mathbf{z}}^l \\
\hat{\mathbf{z}}^{l+1} &= \text{DW-SA}\left(\mathrm{LN}\left(\mathbf{z}^l\right)\right) + \mathbf{z}^l \\
\mathbf{z}^{l+1} &= \mathrm{MLP}\left(\mathrm{LN}\left(\hat{\mathbf{z}}^{l+1}\right)\right) + \hat{\mathbf{z}}^{l+1},
\end{aligned} \tag{5}$$

where $\hat{\mathbf{z}}^l$ and $\mathbf{z}^l$ denote the output features of the W-MSA and DW-SA modules and the MLP module for block l, respectively; W-MSA and LN denote window-based multi-head self-attention and layer norm of Swin-T, respectively.

For $\hat{\mathrm{V}}_\mathrm{D}^{\ \mathrm{F}}$, we utilize ResNet (He et al. 2016) as $\mathrm{Extractor_F}$ to capture spatial features. Both coarse and fine scene spatial extractors are pre-trained on ImageNet-1k (Deng et al. 2009). Given input $\hat{\mathrm{V}}_\mathrm{D}^{\ \mathrm{C}}, \hat{\mathrm{V}}_\mathrm{D}^{\ \mathrm{F}} \in \mathbb{R}^{\mathrm{F}\times\mathrm{W}\times\mathrm{H}\times 3}$, the coarse scene extractor first resizes $\hat{\mathrm{V}}_\mathrm{D}^{\ \mathrm{C}}$ from $\mathrm{F}\times\mathrm{W}\times\mathrm{H}\times 3$ to $\mathrm{F}\times 224\times 224\times 3$, while the fine scene extractor processes $\hat{\mathrm{V}}_\mathrm{D}^{\ \mathrm{F}}$ directly. The final spatial features $\mathrm{F_S} \in \mathbb{R}^{\mathrm{F}\times 5632}$ are as follows:

$$\begin{aligned}
&\mathrm{F_S^C} = \mathrm{Extractor_C}(\hat{\mathrm{V}}_\mathrm{D}^{\ \mathrm{C}}), \mathrm{F_S^F} = \mathrm{Extractor_F}(\hat{\mathrm{V}}_\mathrm{D}^{\ \mathrm{F}}), \\
&\mathrm{F_S} = \mathrm{Concatenate}(\mathrm{V_S^C}, \mathrm{V_S^F}).
\end{aligned} \tag{6}$$

Inconsistency Guided Temporal Module (IGTM)

In this section, we propose a two-stage temporal aggregation integrating cross-time-scale relationships.

Consistency-aware Fusion stage One of the key factors in temporal aggregation is determining the appropriate amount of information that each time segment should load. Recent psychological research on the visual working memory mechanism (Cowan 2001; Schmidt et al. 2004; Dube and Al-Aidroos 2019) has yielded perceptual findings regarding

memory capacity, showing that human visual working memory has a capacity limit when temporarily storing and processing visual information.

Inspired by these findings, we developed a **Visual Memory Capacity Block** that follows two principles:

1. Dynamic allocation of memory threshold in a range.

2. When the level of time inconsistency increases, the memory threshold decreases.

Principle 1 is derived from psychological research (Cowan 2001; Zhang and Luck 2008; Dube and Al-Aidroos 2019; Hajonides et al. 2020), which suggests that the memory capacity of humans is approximately 3-7 objects and dynamically adjusted according to scene complexity. Principle 2 is supported by psychological studies (Schmidt et al. 2004; Dempere-Marco, Melcher, and Deco 2012), which show that increased scene complexity negatively impacts memory performance.

We first quantify the temporal inconsistency levels. Since the video $\mathrm{V_I}$ reflects the level of temporally inconsistent changes in the SR video, we calculate the complexity of $\mathrm{V_I}$ to determine the memory capacity (i.e., threshold). Specifically, we first compute the spatial complexity of each frame and then obtain the overall complexity of the video. As $\mathrm{V_I}$ is based on optical flow, the magnitude of each frame and directional consistency are taken into account:

$$\mathrm{C_I^{ij}} = \alpha \times (\sigma(\mathrm{M}(\mathrm{V_I^{ij}})) + (1-\alpha) \times (\sigma(\mathrm{D}(\mathrm{V_I^{ij}}))), \quad \mathrm{i} = 1, \ldots, \mathrm{N}; \mathrm{j} = 1, \ldots, \mathrm{F}. \tag{7}$$

where N and F are the number of videos and frame count for each $\mathrm{V_I}$, respectively. $\mathrm{M}(\cdot)$ refers to magnitude computation, and $\mathrm{D}(\cdot)$ calculates the direction of the difference between adjacent frames, generating a histogram of the directions, and then the standard deviation σ of the histogram represents the directional consistency. α is a hyperparameter. Then, the complexity of $\mathrm{V_I}$ is calculated as follows:

$$\mathrm{C_I^i} = \mu(\{\mathrm{C_I^{ij}} \| \mathrm{j} = 1, \ldots, \mathrm{F}\}) + \sigma(\{\mathrm{C_I^{ij}} \| \mathrm{j} = 1, \ldots, \mathrm{F}\}), \quad \mathrm{i} = 1, \ldots, \mathrm{N}. \tag{8}$$

where μ and σ refer to mean and standard deviation, respectively. $\mathrm{C_I} = \{\mathrm{C_I^i}\}$ is then normalized to $[0, 1]$. The ablation analysis of α is displayed in Experiments. Then the input features $\mathrm{F_S}$ can be segmented and aggregated by the following visual memory capacity block, as shown in Algorithm 1.

Algorithm 1: Visual Memory Capacity Block

Input: Spatial features $\mathrm{F_S^i} \in \mathbb{R}^{\mathrm{F} \times 5632}$, temporal inconsistency levels $\mathrm{C_I^{ij}} \in \mathbb{R}^{\mathrm{F} \times 1}$ and $\mathrm{C_I^i} \in \mathbb{R}^1$, where $\mathrm{i} = 1, \ldots, \mathrm{N}; \mathrm{j} = 1, \ldots, \mathrm{F}$.

Output: Aggregated features of the first stage: $\mathrm{F_A^i}, \mathrm{i} = 1, \ldots, \mathrm{N}$.

Compute the adaptive memory threshold of SR video $\mathrm{V_D^i}$ as:
$\mathrm{T_D^i} = \tau - \eta \times \frac{\mathrm{C_I^i} - \mathrm{MIN}(\{\mathrm{C_I^i}\})}{\mathrm{MAX}(\{\mathrm{C_I^i}\}) - \mathrm{MIN}(\{\mathrm{C_I^i}\})}$.
\# The adaptive $\mathrm{T_D^i}$ uses $\tau = 5$ and $\eta = 4$.
Adaptively segment input features $\mathrm{F_S^i}$:
Set of segments:$\mathrm{S} = \varnothing$;
Current segment:$\mathrm{S_C} = \varnothing$;
for $\mathrm{j} \in [1, \mathrm{F}]$ **do**
  Current complexity:$\mathrm{C_C} += \mathrm{C_I^{ij}}$
  Add $\mathrm{F_S^{ij}}$ to $\mathrm{S_C}$: $\mathrm{S_C}.\mathrm{append}(\mathrm{F_S^{ij}})$
  if $\mathrm{C_C} >= \mathrm{T_D^i}$ **then**
    $\mathrm{S}.\mathrm{append}(\mathrm{S_C})$
    $\mathrm{S_C} = \varnothing$
    $\mathrm{C_C} = 0$
  end if
end for
for $\mathrm{k} \in \mathrm{N_S}$ **do**
  $\mathrm{F_A^{ik}} = \mathrm{Concatenate}(\mathrm{MEAN}(\mathrm{S_k}) + \mathrm{STD}(\mathrm{S_k}))$
end for
\#Where $\mathrm{N_S}$ is the number of segments in S.

Subsequently, we model the temporal relationships of the first-stage features $\mathrm{F_A} \in \{\mathrm{F_A^i}\}$. Recognizing that Long Short-Term Memory models (Hochreiter and Schmidhuber 1997; Chung et al. 2014) are insufficient for capturing complex temporal dependencies, we propose the **Temporal Modeling** process as follows:

First, we calculate the sparse adjacency matrix of input features $\mathrm{Adj}(\mathrm{F_A})$. Each node feature $\mathrm{h_i}$ of $\mathrm{F_A}$ is then transformed using a learnable weight matrix W to increase the representational capacity:

$$\mathrm{W_h(i)} = \mathrm{Wh_i}, \tag{9}$$

where W is the weight matrix shared across all nodes. Then, the attention coefficients $\mathrm{e_{ij}}$ between node i and its neighbor j are computed by concatenating the transformed features of the two nodes, followed by applying a shared attention mechanism:

$$\mathrm{e_{ij}} = \mathrm{LeakyReLU}(\mathrm{a^T}[\mathrm{W_h(i)} \| \mathrm{W_h(j)}]), \tag{10}$$

where a is the learnable attention vector and $\|$ denotes concatenation. Then filter $\mathrm{e} = \{\mathrm{e_{ij}}\}$ based on the positive elements of $\mathrm{Adj}(\mathrm{F_A})$:

$$\mathrm{e_{ij}} = \begin{cases} \mathrm{e_{ij}} & \mathrm{Adj(F_A)_{ij}} > 0 \\ 0 & \mathrm{Adj(F_A)_{ij}} \leq 0 \end{cases}$$

The attention coefficients are normalized using the softmax function:

$$\alpha_{\mathrm{ij}} = \frac{\exp(\mathrm{e_{ij}})}{\sum_{\mathrm{k} \in \mathcal{N}(\mathrm{i})} \exp(\mathrm{e_{ik}})}, \tag{11}$$

where $\mathcal{N}(\mathrm{i})$ represents the neighbors of node i. Finally, the new feature $\mathrm{h_i'}$ of node i is computed by aggregating its neighbors' features weighted by the attention coefficients:

$$\mathrm{h_i'} = \phi\left(\sum_{\mathrm{j} \in \mathcal{N}(\mathrm{i})} \alpha_{\mathrm{ij}} \mathrm{W_h(j)}\right), \tag{12}$$

where ϕ is a non-linear activation function. And the features after time modeling are obtained by the Gated Recurrent Unit (GRU) (Chung et al. 2014):

$$\mathrm{F_{S_1}} = \mathrm{GRU}(\{\mathrm{h_i'}\}). \tag{13}$$

These processes together make up one temporal modeling process. $\mathrm{F_{S_1}}$ is then regressed to score $\mathrm{S_1}$, and used for the second-stage temporal aggregation.

Datasets	SFD				MFD				Combined-VSR			
Methods	SRCC↑	PLCC↑	KRCC↑	RMSE↓	SRCC↑	PLCC↑	KRCC↑	RMSE↓	SRCC↑	PLCC↑	KRCC↑	RMSE↓
PSNR	0.654	0.661	0.476	0.197	0.639	0.648	0.459	0.202	0.645	0.655	0.468	0.200
SSIM (Wang et al. 2004)	0.709	0.719	0.534	0.185	0.693	0.701	0.519	0.190	0.696	0.710	0.525	0.189
VIF (Sheikh and Bovik 2006)	0.746	0.755	0.579	0.164	0.742	0.749	0.571	0.166	0.746	0.753	0.579	0.165
SpEED-QA (Bampis et al. 2017)	0.507	0.520	0.309	0.243	0.501	0.514	0.303	0.247	0.504	0.516	0.307	0.245
IGTS (Tang et al. 2019)	0.535	0.552	0.338	0.232	0.530	0.547	0.334	0.235	0.533	0.550	0.337	0.234
VSFA (Li, Jiang, and Jiang 2019)	0.813	0.819	0.636	0.150	0.796	0.801	0.619	0.158	0.808	0.812	0.630	0.152
DeepSRQ (Zhou et al. 2020)	0.671	0.661	0.480	0.192	0.651	0.643	0.462	0.201	0.667	0.656	0.474	0.195
VMAF (Orduna et al. 2020)	0.712	0.729	0.545	0.182	0.706	0.713	0.528	0.188	0.710	0.723	0.539	0.184
DR-IQA (Zheng et al. 2021)	0.717	0.723	0.542	0.179	0.701	0.711	0.532	0.188	0.707	0.716	0.537	0.184
VIDEVAL (Tu et al. 2021)	0.754	0.759	0.583	0.164	0.731	0.740	0.569	0.171	0.744	0.749	0.577	0.167
DISQ (Zhao et al. 2022)	0.665	0.661	0.475	0.198	0.636	0.641	0.459	0.207	0.642	0.650	0.465	0.202
SRIF (Zhou and Wang 2022)	0.751	0.758	0.582	0.167	0.739	0.744	0.570	0.171	0.743	0.751	0.576	0.169
GSTVQA (Chen et al. 2022)	0.839	0.837	0.655	0.141	0.816	0.815	0.633	0.153	0.828	0.825	0.645	0.147
STF (Zhou et al. 2023)	0.776	0.783	0.634	0.158	0.754	0.758	0.583	0.169	0.763	0.769	0.594	0.163
STI-VQA (Zhu et al. 2023)	0.835	0.842	0.658	0.139	0.814	0.819	0.636	0.152	0.823	0.829	0.648	0.147
2Bi-VQA (Telili et al. 2023)	0.820	0.838	0.628	0.125	0.774	0.809	0.569	0.140	0.777	0.810	0.585	0.665
FAST-VQA (Wu et al. 2022)	0.852	0.861	0.661	0.130	0.836	0.851	0.639	0.134	0.845	0.856	0.651	0.132
MBVQA (Wen et al. 2024)	0.892	0.901	0.717	0.106	0.795	0.827	0.606	0.124	0.840	0.853	0.644	0.127
VSR-QAD(Zhou et al. 2024)	0.884	0.889	0.706	0.115	0.846	0.846	0.672	0.134	0.860	0.868	0.687	0.125
ReLaX-VQA (Wang, Katsenou, and Bull 2025)	0.937	0.949	0.785	0.100	0.915	0.925	0.747	0.098	0.924	0.936	0.782	0.091
TIG-SVQA	**0.950**	**0.951**	**0.803**	**0.093**	**0.942**	**0.943**	**0.790**	**0.091**	**0.939**	**0.942**	**0.794**	**0.083**

Table 1: Performance comparison of our methods against competing IQA/VQA methods on both Single-Frame SR videos and Multi-Frame SR videos. The best and second-best performances are highlighted in bold and underlined, respectively.

Methods	SRCC↑	PLCC↑	KRCC↑	RMSE↓
w/o Guidance in IHSM	0.891	0.909	0.716	0.116
w/o Guidance in IGTM	0.908	0.921	0.736	0.095
w/o both Guidance	0.878	0.901	0.707	0.107
Coarse Branch only in IHSM	0.825	0.886	0.646	0.113
Fine Branch only in IHSM	0.926	0.927	0.771	0.106
w/o DW-SA-T block	0.891	0.909	0.716	0.116
Proposed TIG-SVQA	**0.939**	**0.942**	**0.794**	**0.083**

Table 2: Ablation experiments on each component of TIG-SVQA on the Combined-VSR dataset. Note that "w/o Guidance in IGTM" is to replace the visual memory capacity block with fixed temporal segments.

Model	Flops/G ↓	Paras/M ↓	SRCC↑
DISQ (Zhao et al. 2022)	606.69	76.18	0.642
FAST-VQA (Wu et al. 2022)	**70.90**	27.55	0.845
STI-VQA (Zhu et al. 2023)	103087.70	89.37	0.823
MBVQA (Wen et al. 2024)	2149.90	93.23	0.840
VSR-QAD (Zhou et al. 2024)	678.95	**23.74**	0.860
Proposed TIG-SVQA	171.63	24.96	**0.939**

Table 3: Comparisons regarding model complexity. Note that the number of parameters of methods that contain several modules is the sum of each module. The best and second-best performances are highlighted in bold and underlined, respectively.

Informative Filtering stage In this stage, we filter out the most informative features to facilitate temporal relationship modeling. Specifically, self-attention is computed over $\mathrm{F_{S_1}}$, and the top K features with the highest attention scores are selected as key features:

$$\mathrm{W_S} = \mathrm{SA}(\mathrm{F_{S_1}}), \mathrm{F_{S_2}} = \mathrm{F_{S_1}}[\mathrm{Top_K}(\mathrm{W_S})], \tag{14}$$

where SA represents self-attention. $\mathrm{W_S}$ is the attention weight. The $\mathrm{F_{S_2}}$ is then processed by another temporal modeling process, which regresses the second-stage quality score $\mathrm{S_2}$. The final quality prediction is computed as:

$$\mathrm{S} = \gamma \times \mathrm{S_1} + (1 - \gamma) \times \mathrm{S_2}, \tag{15}$$

where γ is a hyperparameter.

Experiments

Experimental Setups

Benchmark Datasets. We evaluate TIG-SVQA on both single-frame and multi-frame super-resolution video quality datasets, i.e., SFD and MFD. The SFD (Zhou et al. 2024) contains 1,193 videos generated from 120 reference sequences, downsampled by factors of ×2, ×4, and ×8, and then upscaled using five different single-frame SR methods. In contrast, the MFD (Zhou et al. 2024) comprises 1,067 videos produced by five state-of-the-art multi-frame SR algorithms. The Combined-VSR dataset includes all 2,260 videos, spanning 10 recent SR methods across both single-frame and multi-frame categories.

Competing Methods. As shown in Table 1, we compare our method with 8 image quality assessment (IQA) (i.e., SSIM, VIF, SRIF, IGTS, DISQ, DR-IQA, STF, DeepSRQ), and 10 video quality assessment (VQA) methods (i.e., VIDEVAL, VMAF, SpEED-QA, VSRQAD, VSFA, GSTVQA, STI-VQA, 2Bi-VQA, FAST-VQA, MBVQA, ReLaX-VQA). Among them, SRIF, IGTS, DISQ, STF, and DeepSRQ are SR IQA methods; and VSR-QAD is the latest SR VQA method.

Performance Criteria. We evaluate performance using Spearman rank-order correlation coefficient (SRCC), Kendall rank-order correlation coefficient (KRCC), Pearson linear correlation coefficient (PLCC), and root mean square error (RMSE). For each dataset, we follow the protocols of (Zhou et al. 2024): randomly splitting videos into 70% training, 10% validation, and 20% testing.

Implementation Details. Models were trained for 100 epochs on a 16GB NVIDIA RTX 3080 Ti with PyTorch 1.7.1. The Adam optimizer was used with an initial learning rate of 10^{-5}, decaying by 0.8 every 10 epochs. The batch

size was 16, with no weight decay. The loss combined SRCC and MSE to leverage both ranking and regression.

The ablation studies on individual components and hyperparameters, as well as the analysis of model complexity, are presented in the following sections. Additional ablation results, including those on the adaptive memory threshold, loss function, the DW-SA Transformer within IHSM, the rationale of IGTM design, and the model's generalization to real-world scenarios, are provided in the extended version (Li et al. 2025).

Performance on SFD, MFD, and Combined-VSR

Table 1 presents a comprehensive comparison of the latest IQA/VQA methods on single-frame and multi-frame SR video quality datasets. Among the listed methods, PSNR and SSIM are classical handcrafted metrics, relying on pixel-wise differences without learning, while VIF and SpEED-QA are also handcrafted metrics utilizing signal fidelity and statistical priors. In contrast, methods such as DeepSRQ, FAST-VQA, MBVQA, ReLaX-VQA, and VSR-QAD are learning-based approaches. Notably, SRIF, IGTS, DISQ, STF, and DeepSRQ are originally proposed for SR image quality assessment, whereas VSR-QAD is the latest VQA method designed specifically for SR videos.

Compared to handcrafted metrics like VIF and SpEED-QA, TIG-SVQA achieves substantial improvements in correlation metrics and error reduction, particularly on the more challenging MFD subset. Even when compared to the latest traditional VQA methods, including MBVQA and ReLaX-VQA, TIG-SVQA shows superior results. These findings highlight the necessity for dedicated temporal inconsistency guidance for learning SR video distortions.

Ablation Study

The ablation of each component in TIG-SVQA highlights the individual contributions, as shown in Table 2. **w/o Guidance in IHSM** refers to using SR videos as input without weighting with temporal inconsistency information in the IHSM. **w/o Guidance in IGTM** refers to replacing the visual memory capacity block in the temporal module with fixed temporal segments (i.e., one batch has 16 frames). **w/o both Guidance** refers to not using temporal inconsistency guidance in both IHSM and IGTM. **Coarse/Fine Branch only in IHSM** refers to only using the coarse/fine scene spatial extractor. And **w/o DW-SA-T block** refers to utilizing Swin-T blocks rather than the proposed DW-SA-T blocks in the IHSM. The model's performance drops notably without the DW-SA Transformer blocks and inconsistency guidance, highlighting their critical role in perceptual prediction. The coarse/fine spatial branch alone shows a marked decline, emphasizing the need to combine both granularity features.

Model Complexity Analysis and Parameter Test

Table 3 compares state-of-the-art VQA methods in terms of **FLOPs and parameter count**. FLOPS is shown in GigaFLOPS (G), while the parameter count is shown in Mbyte (M). We assume that the resolution of the test video is 1080P, the duration is 6 seconds, and the number of video frames is 150 frames. It can be observed that TIG-SVQA has the best correlation performance and the second-best regarding FLOPs as well as parameters.

Hyper-Param-α	SRCC↑	PLCC↑	KRCC↑	RMSE↓
0.1	0.931	0.935	0.782	0.101
0.3	0.933	0.941	0.779	0.100
0.5	**0.939**	**0.942**	**0.794**	**0.083**
0.7	0.930	0.935	0.784	0.087
0.9	0.929	0.936	0.782	0.103

Table 4: Ablation on the hyper-parameter α, which determines the weight of the magnitude complexity and complexity of direction consistency when computing the level of temporal inconsistency for each frame.

Hyper-Param-γ	SRCC↑	PLCC↑	KRCC↑	RMSE↓
0.0	0.919	0.931	0.764	0.106
0.2	0.926	0.935	0.773	0.094
0.4	0.932	0.938	0.783	0.091
0.5	**0.939**	**0.942**	**0.794**	**0.083**
0.6	0.928	0.936	0.779	0.090
0.8	0.918	0.927	0.764	0.090
1.0	0.910	0.920	0.748	0.092

Table 5: Ablation on the hyper-parameter γ, which determines the proportion of predicted scores S_1 and S_2 at cross-time-scales in the final predicted score S.

Table 4 presents an ablation on **hyperparameter** α. The model performs best at $\alpha = 0.5$. Performance decreases at the extremes ($\alpha = 0.1$ and $\alpha = 0.9$), indicating that both magnitude and direction are key optical flow cues (Gujjunoori, Oruganti, and Pais 2022). The **hyperparameter** γ is used to weight the prediction quality scores S_1 and S_2 of the two stages. As shown in Table 5, the results show that two-stage temporal modeling outperforms single-stage, reflecting hierarchical human temporal perception (Hasson et al. 2008).

Conclusion

This paper presents TIG-SVQA, a novel framework for super-resolution video quality assessment. We begin by quantifying temporal inconsistency and demonstrating its strong correlation with human perception. Using this insight, TIG-SVQA improves the accuracy of perceptual prediction under the guidance of temporal inconsistency information. The proposed method first introduces an Inconsistency Highlighted Spatial Module that extracts coarse- and fine-grained inconsistency features, incorporating a newly designed DW-SA-T block. Then, we propose an Inconsistency Guided Temporal Module, which consists of consistency-aware fusion and informative filtering stages. A visual memory capacity block adaptively allocates temporal segments in the first stage, while the filtering process further emphasizes quality-oriented features in the second stage. Extensive experiments on both single-frame and multi-frame SR video quality datasets demonstrate the superior performance of TIG-SVQA.

Acknowledgments

This work was in part supported by the National Natural Science Foundation of China under Grant 62371017. This work was done when Yixiao Li was an academic visitor at Cardiff University.

References

Bampis, C. G.; Gupta, P.; Soundararajan, R.; and Bovik, A. C. 2017. SpEED-QA: Spatial Efficient Entropic Differencing for Image and Video Quality. *IEEE Signal Processing Letters*, 24(9): 1333–1337.

Banitalebi-Dehkordi, M.; Ebrahimi-Moghadam, A.; Khademi, M.; and Hadizadeh, H. 2020. No-Reference Video Quality Assessment Based on Visual Memory Modeling. *IEEE Transactions on Broadcasting*, 66(3): 676–689.

Bi, X.; He, X.; Xiong, S.; Zhao, Z.; Chen, H.; and Sheriff, R. E. 2024. Blind video quality assessment based on Spatio-Temporal Feature Resolver. *Neurocomputing*, 574: 127249.

Cao, Y.; Sun, W.; Zhang, W.; Sun, Y.; Jia, Z.; Zhu, Y.; Min, X.; and Zhai, G. 2024. SR-VQA: Super-Resolution Video Quality Assessment Model. In *European Conference on Computer Vision Workshop*, 144–159.

Chan, K. C.; Zhou, S.; Xu, X.; and Loy, C. C. 2022. BasicVSR++: Improving video super-resolution with enhanced propagation and alignment. In *IEEE CVPR*.

Chen, B.; Zhu, L.; Li, G.; Lu, F.; Fan, H.; and Wang, S. 2022. Learning Generalized Spatial-Temporal Deep Feature Representation for No-Reference Video Quality Assessment. *IEEE TCSVT*, 32(4): 1903–1916.

Chung, J.; Çaglar Gülçehre; Cho, K.; and Bengio, Y. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *ArXiv*, abs/1412.3555.

Cowan, N. 2001. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24: 87 – 114.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable Convolutional Networks. In *IEEE ICCV*, 764–773.

Dempere-Marco, L.; Melcher, D.; and Deco, G. 2012. Effective Visual Working Memory Capacity: An Emergent Effect from the Neural Dynamics in an Attractor Network. *PLoS ONE*, 7.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *IEEE CVPR*, 248–255.

Dube, B.; and Al-Aidroos, N. 2019. Distinct prioritization of visual working memory representations for search and for recall. *Attention, Perception, & Psychophysics*, 81: 1253 – 1261.

Ebenezer, J. P.; Shang, Z.; Wu, Y.; Wei, H.; Sethuraman, S.; and Bovik, A. C. 2020. ChipQA: No-Reference Video Quality Prediction via Space-Time Chips. *IEEE TIP*, 8059–8074.

Gonzalez, R.; Woods, R.; and Masters, B. 2009. Digital Image Processing, Third Edition. *Journal of Biomedical Optics*, 14: 029901.

Gujjunoori, S.; Oruganti, M.; and Pais, A. R. 2022. Enhanced optical flow-based full reference video quality assessment algorithm. *Multimedia Tools and Applications*, 81(27): 39491–39505.

Hajonides, J. E.; van Ede, F.; Stokes, M. G.; and Nobre, A. C. 2020. Comparing the prioritization of items and feature-dimensions in visual working memory. *Journal of Vision*, 20.

Hasson, U.; Yang, E.; Vallines, I.; Heeger, D. J.; and Rubin, N. 2008. A hierarchy of temporal receptive windows in human cortex. *Journal of neuroscience*, 28(10): 2539–2550.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *IEEE CVPR*, 770–778.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Computation*, 9(8): 1735–1780.

Hosu, V.; Hahn, F.; Jenadeleh, M.; Lin, H.; Men, H.; Szirányi, T.; Li, S.; and Saupe, D. 2017. The Konstanz natural video database (KoNViD-1k). In *IEEE International Conference on Quality of Multimedia Experience*, 1–6.

Imani, H.; Islam, M. B.; and Wong, L.-K. 2022. A New Dataset and Transformer for Stereoscopic Video Super-Resolution. *IEEE CVPRW*, 705–714.

Kirillova, A.; Lyapustin, E.; Antsiferova, A.; and Vatolin, D. 2021. ERQA: Edge-restoration quality assessment for video super-resolution. *arXiv preprint arXiv:2110.09992*.

Korhonen, J. 2019. Two-Level Approach for No-Reference Consumer Video Quality Assessment. *IEEE TIP*, 28(12): 5923–5938.

Li, B.; Zhang, W.; Tian, M.; Zhai, G.; and Wang, X. 2022. Blindly Assess Quality of In-the-Wild Videos via Quality-aware Pre-training and Motion Perception. *IEEE TCSVT*, 32(9): 5944–5958.

Li, D.; Jiang, T.; and Jiang, M. 2019. Quality Assessment of In-the-Wild Videos. In *ACM International Conference on Multimedia*, 2351–2359.

Li, Y.; Yang, X.; Liu, W.; Jin, X.; Jia, X.; Lai, Y.; Rosin, P. L.; Liu, H.; and Zhou, W. 2025. Temporal Inconsistency Guidance for Super-resolution Video Quality Assessment. arXiv:2412.18933.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *ICCV*.

Orduna, M.; Díaz, C.; Muñoz, L.; Pérez, P.; Benito, I.; and García, N. 2020. Video Multimethod Assessment Fusion (VMAF) on 360VR Contents. *IEEE Transactions on Consumer Electronics*, 66(1): 22–31.

Postle, B. R. 2015. The cognitive neuroscience of visual short-term memory. *Current Opinion in Behavioral Sciences*, 1: 40–46.

Saad, M. A.; Bovik, A. C.; and Charrier, C. 2014. Blind Prediction of Natural Video Quality. *IEEE TIP*, 1352–1365.

Santangelo, V.; and Macaluso, E. 2013. Visual Salience Improves Spatial Working Memory via Enhanced Parieto-Temporal Functional Connectivity. *The Journal of Neuroscience*, 33: 4110 – 4117.

Schmidt, M.; Jin, S.-W.; Gray, A. M.; Beis, D.; Pham, T.; Frantz, G. D.; Palmieri, S.; Hillan, K. S. F.; Stainier, D. Y. R.; de Sauvage, F. J.; and Ye, W. 2004. Capacity limit of visual short-term memory in human posterior parietal cortex. *Nature*, 428: 751–754.

Sheikh, H.; and Bovik, A. 2006. Image information and visual quality. *IEEE TIP*, 15(2): 430–444.

Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A. P.; Bishop, R.; Rueckert, D.; and Wang, Z. 2016. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In *IEEE CVPR*, 1874–1883.

Sigala, N.; Kaldy, Z.; and Reynolds, G. D. 2022. Editorial: The cognitive neuroscience of visual working memory, Volume II. *Frontiers in Systems Neuroscience*, 16.

Sinno, Z.; and Bovik, A. C. 2019. Large-Scale Study of Perceptual Video Quality. *IEEE TIP*, 28(2): 612–627.

Sun, W.; Min, X.; Lu, W.; and Zhai, G. 2022. A Deep Learning Based No-Reference Quality Assessment Model for UGC Videos. In *ACM International Conference on Multimedia*, 856–865.

Tang, L.; Sun, K.; Liu, L.; Wang, G.; and Liu, Y. 2019. A reduced-reference quality assessment metric for super-resolution reconstructed images with information gain and texture similarity. *Signal Processing: Image Communication*, 79: 32–39.

Telili, A.; Fezza, S. A.; Hamidouche, W.; and Brachemi Meftah, H. F. Z. 2023. 2BiVQA: Double Bi-LSTM-based Video Quality Assessment of UGC Videos. *ACM Trans. Multimedia Comput. Commun. Appl.*, 20(4).

Tu, Z.; Chen, C.-J.; Wang, Y.; Birkbeck, N.; Adsumilli, B.; and Bovik, A. C. 2021. Video Quality Assessment of User Generated Content: A Benchmark Study and a New Model. In *IEEE International Conference on Image Processing*.

Wang, X.; Katsenou, A.; and Bull, D. 2025. Relax-VQA: Residual fragment and layer stack extraction for enhancing video quality assessment. *arXiv preprint arXiv:2407.11496*.

Wang, Z.; Bovik, A.; Sheikh, H.; and Simoncelli, E. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4): 600–612.

Wen, W.; Li, M.; Zhang, Y.; Liao, Y.; Li, J.; Zhang, L.; and Ma, K. 2024. Modular Blind Video Quality Assessment. In *IEEE CVPR*, 2763–2772.

Wu, H.; Chen, C.; Hou, J.; Liao, L.; Wang, A.; Sun, W.; Yan, Q.; and Lin, W. 2022. FAST-VQA: Efficient end-to-end video quality assessment with fragment sampling. In *European conference on computer vision*, 538–554. Springer.

Xu, J.; Li, J.; Zhou, X.; Zhou, W.; Wang, B.; and Chen, Z. 2021. Perceptual quality assessment of internet videos. In *ACM International Conference on Multimedia*, 1248–1257.

You, J.; and Lin, Y. 2022. Efficient Transformer with Locally Shared Attention for Video Quality Assessment. In *IEEE International Conference on Image Processing*, 356–360.

Zhang, H.; Liu, D.; and Xiong, Z. 2020. Is There Tradeoff between Spatial and Temporal in Video Super-Resolution? *arXiv preprint arXiv:2003.06141*.

Zhang, W.; and Luck, S. J. 2008. Discrete fixed-resolution representations in visual working memory. *Nature*, 453: 233–235.

Zhao, T.; Lin, Y.; Xu, Y.; Chen, W.; and Wang, Z. 2022. Learning-Based Quality Assessment for Image Super-Resolution. *IEEE TMM*, 24: 3570–3581.

Zheng, H.; Yang, H.; Fu, J.; Zha, Z.-J.; and Luo, J. 2021. Learning Conditional Knowledge Distillation for Degraded-Reference Image Quality Assessment. In *IEEE ICCV*, 10242–10251.

Zhou, F.; Sheng, W.; Lu, Z.; Kang, B.; Chen, M.; and Qiu, G. 2023. Super-resolution image visual quality assessment based on structure–texture features. *Signal Processing: Image Communication*, 117: 117025.

Zhou, F.; Sheng, W.; Lu, Z.; and Qiu, G. 2024. A Database and Model for the Visual Quality Assessment of Super-Resolution Videos. *IEEE Transactions on Broadcasting*, 70(2): 516–532.

Zhou, W.; Jiang, Q.; Wang, Y.; Chen, Z.; and Li, W. 2020. Blind quality assessment for image superresolution using deep two-stream convolutional networks. *Information Sciences*, 528: 205–218.

Zhou, W.; and Wang, Z. 2022. Quality Assessment of Image Super-Resolution: Balancing Deterministic and Statistical Fidelity. In *ACM International Conference on Multimedia*, 934–942. Association for Computing Machinery.

Zhu, H.; Chen, B.; Zhu, L.; and Wang, S. 2023. Learning Spatiotemporal Interactions for User-Generated Video Quality Assessment. *IEEE TCSVT*, 33(3): 1031–1042.

Zhu, X.; Hu, H.; Lin, S.; and Dai, J. 2019. Deformable ConvNets V2: More Deformable, Better Results. In *IEEE CVPR*, 9300–9308.