QMamba: On First Exploration of Vision Mamba for Image Quality Assessment

Fengbin Guan [1] **Xin Li** [1] **Zihao Yu** [1] **Yiting Lu** [1] **Zhibo Chen** [1]

Abstract

In this work, we take the first exploration of the recently popular foundation model, *i.e.*, State Space Model/Mamba, in image quality assessment (IQA), aiming at observing and excavating the perception potential in vision Mamba. A series of works on Mamba has shown its significant potential in various fields, *e.g.*, segmentation and classification. However, the perception capability of Mamba remains under-explored. Consequently, we propose QMamba by revisiting and adapting the Mamba model for three crucial IQA tasks, *i.e.*, task-specific, universal, and transferable IQA, which reveals its clear advantages over existing foundational models, *e.g.*, Swin Transformer, ViT, and CNNs, in terms of perception and computational cost. To improve the transferability of QMamba, we propose the StylePrompt tuning paradigm, where lightweight mean and variance prompts are injected to assist task-adaptive transfer learning of pre-trained QMamba for different downstream IQA tasks. Compared with existing prompt tuning strategies, our StylePrompt enables better perceptual transfer with lower computational cost. Extensive experiments on multiple synthetic, authentic IQA datasets, and cross IQA datasets demonstrate the effectiveness of our proposed QMamba. The code will be available at: https://github.com/bingo-G/QMamba.git

1. Introduction

Image Quality Assessment (IQA) aims to measure the subjective quality of images aligned with human perception, which has been applied in various visual fields, including visual acquisition, transmission, AIGC (Li et al., 2023a; Wang et al., 2023a), and UGC creation (Tu et al., 2021; Lu et al., 2024), etc. Establishing a great IQA metric is necessary to provide the right optimization direction tailored for image processing techniques, *i.e.*, compression (Li et al., 2021; Wu et al., 2021; Yu et al., 2024b), enhancement (Li et al., 2023b; Fei et al., 2023), and ensure the perceptual quality of images. Early works on IQA have been achieved by leveraging the natural scene statistics in a hand-crafted manner (Mittal et al., 2012a;b). With the advancements of deep neural networks (DNNs), learning-based IQA metrics (Li et al., 2023c; Liu et al., 2022) have demonstrated significant potential for low-level perception, which can be roughly categorized into two types based on the pre-trained backbones: CNN-based and Transformer-based methods.

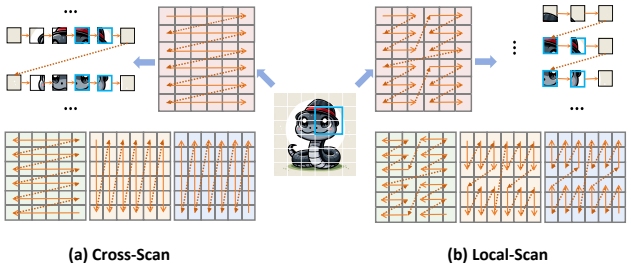

Figure 1. Scanning Methodology Illustration. (a) VMamba's scanning method(Liu et al., 2024c) flattens 2D data into 1D, impairing connectivity by distancing adjacent tokens. (b) Local scanning method scans within and across windows, placing semantically similar and distortion-related tokens closer, as shown in the blue boxes.

Although the impressive progress, learning-based IQA is susceptible to inherent limitations of existing pre-trained backbones: (i) the CNNs are skilled at learning local translation-invariant features from images while lacking enough long-range dependency modeling capability, hindering the global quality perception. (ii) The emergence of Vision Transformers presents a great solution to model long-range dependency effectively by leveraging attention mechanisms. However, the quadratic complexity of self-attention operations poses unaffordable computational costs, especially for large-scale image quality assessment. Recently, an innovative foundation model, the State Space Model, particularly its implementation, *i.e.*, Mamba (Gu & Dao, 2023) has shown considerable potential in various fields for balancing the computational costs and performances,

[1]University of Science and Technology of China, Hefei, China. Correspondence to: Xin Li <xin.li@ustc.edu.cn>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

e.g., segmentation(Xing et al., 2024; Ma et al., 2024) and classification(Liu et al., 2024c; Zhu et al., 2024). This raises one interesting question, *i.e.*, "whether Mamba can surpass existing backbones on low-level visual perception", which is still under-explored.

To answer this question, in this work, we initiate a new exploration of the Mamba model within the field of IQA and introduce the QMamba, a newly developed IQA framework designed to address three key facets of IQA: task-specific, universal, and transferable image quality assessment. Notably, since the lack of enough IQA dataset, learning-based IQA metrics entail the pre-trained backbone for perception knowledge extraction. Meanwhile, VMamba(Liu et al., 2024c), as the most representative framework, has achieved excellent performance on high-level tasks by employing horizontal and vertical scanning strategies. However, merely excavating the global perception knowledge is not optimal for IQA, since most artifacts affecting the image quality are related to local textures. Consequently, inspired by LocalMamba(Huang et al., 2024), we have adopted a local window scanning method, which significantly enhances ability of Mamba to perceive local distortions, thereby demonstrating superior performance. We detail the architecture of QMamba and examine different scanning methods to clarify how the model interacts with and processes image data. Our extensive analysis confirms that QMamba surpasses traditional foundational models, showing superior perceptual accuracy and greater computational efficiency.

Moreover, we have explored the perceptual transferability of Mamba across different datasets, *i.e.*, different contents, and degradations. We can find that the Mamba-based IQA metric still suffers from severe performance drop when they encounter large domain shifts between synthetic distortions(Sheikh et al., 2006; Lin et al., 2019), authentic distortions(Ghadiyaram & Bovik, 2015; Hosu et al., 2020), and Artificial Intelligence-Generated Content (AIGC)(Li et al., 2023a; Wang et al., 2023a) distortions, etc. To further amplify the transferability of QMamba across a range of IQA applications, we introduce a simple but effective tuning strategy called StylePrompt. This is based on the finding that the domain shifts in IQA tend to correlate with their feature statistics/style (Lu et al., 2022), such as mean and variance. Concretely, our StylePrompt aims to adaptively adjust the mean and variance of pre-trained QMamba towards the target IQA tasks by setting a group of light-weight learnable $1 \times 1 \times C$ parameters. Extensive experiments have shown that our StylePrompt greatly facilitates the task-adaptive learning of QMamba, enabling efficient knowledge transfer across diverse IQA tasks with fewer parameter costs.

The main contributions of this paper are summarized as follows:

- We embark on a novel exploration of the Mamba model within image quality assessment, and propose the QMamba, a powerful IQA metric for three critical tasks of IQA: task-specific, universal, and transferable image quality assessment. This exploration has demonstrated the superior potential of Mamba for subjective perception, advancing the development of IQA.
- To improve the perception transferability of the QMamba, we introduce a simple but effective tuning strategy, *i.e.*, StylePrompt. This strategy enables the efficient knowledge transfer of pre-trained QMamba for downstream IQA tasks, while only tuning fewer learnable parameters to adjust the statistics of perception features.
- Extensive experiments have shown that our QMamba has consistently achieved state-of-the-art results on various prominent IQA datasets compared with existing IQA methods, thereby validating the efficacy of the Mamba model in quality assessment. Moreover, our StylePrompt achieves nearly equivalent performance to full model tuning while utilizing only 4% of the whole parameters, demonstrating its effectiveness in perception knowledge transfer scenarios.

2. Related Work

2.1. Blind Image Quality Assessment (BIQA)

Early BIQA methods relied heavily on manually designed features for quality score regression (Mittal et al., 2012a;b; Venkatanath et al., 2015; Saad et al., 2012; Min et al., 2018). However, these handcrafted features were insufficient for addressing the complexity of BIQA tasks. With the advent of deep learning, network architectures capable of powerful feature extraction significantly improved quality assessment tasks, with Convolutional Neural Networks (CNNs) and Vision Transformers being the most prevalent.

CNN-based BIQA. CNNs have demonstrated robust feature extraction capabilities, leading to their widespread adoption in BIQA tasks. Early works like CNNIQA (Kang et al., 2014) used convolutional models for feature learning and quality regression, substantially outperforming handcrafted features. DBCNN (Zhang et al., 2020) introduced a dual-stream network to address synthetic and authentic distortions separately, integrating these insights for better quality prediction. NIMA (Talebi & Milanfar, 2018) and PQR (Zeng et al., 2017) leveraged pre-trained models on ImageNet for quality score prediction, enhancing accuracy through well-established neural architectures. MetaIQA (Zhu et al., 2020) employed meta-learning to adapt to unknown distortions by learning shared priors for various distortion types, while HyperIQA (Su et al., 2020) used a hypernetwork to adaptively establish perceptual rules, improving generalization. Despite these advancements(Saha

et al., 2023; Zhao et al., 2023; Chen et al., 2024a), CNNs' local bias limits their ability to fully exploit both global and local information in BIQA tasks.

Transformer-based BIQA. Transformers offer superior global modeling capabilities compared to CNNs. TReS (Golestaneh et al., 2022) addressed CNNs' local bias by capturing local structural information with CNNs and then using Transformers for sequential feature extraction. MUSIQ (Ke et al., 2021) designed a multi-scale image Transformer architecture capable of handling images with varying sizes and aspect ratios. DEIQT (Qin et al., 2023) leveraged a Transformer-based BIQA architecture with attention mechanisms to align with human perception, enhancing model performance and reducing prediction uncertainty. However, the quadratic complexity of Transformers presents a challenge, highlighting the need for architectures with linear complexity capable of global modeling for BIQA tasks. In addition to the above methods, many recent works (Zhang et al., 2023; Shin et al., 2024; Yu et al., 2024a) have also adopted Transformer-based designs and achieved competitive performance, yet they still face limitations due to high computational complexity.

2.2. State Space Models (SSMs)

State space models, known for their linear complexity in capturing long-range dependencies, have been integrated into deep learning architectures. The Structured State Space model (S4) (Gu et al., 2021) was a pioneer in deep state space modeling for remote dependency modeling. Subsequent advancements (Gu et al., 2020; Smith et al., 2022; Fu et al., 2022) further propelled the development of state space models. Mamba (Gu & Dao, 2023), by integrating selection mechanisms and hardware-aware algorithms, has shown effective long-range modeling capabilities with linear complexity growth.

Initially focused on NLP tasks, Mamba has rapidly expanded into other domains. Vim (Zhu et al., 2024) introduced a bidirectional SSM block for visual representation learning, achieving performance comparable to ViT (Dosovitskiy et al., 2020). VMamba (Liu et al., 2024c) introduced a cross-scan module to traverse spatial domains and convert non-causal visual images into ordered block sequences, maintaining linear complexity while retaining global receptive fields. LocalMamba (Huang et al., 2024) employed a window-based scanning approach to integrate local inductive biases, enhancing the visual Mamba model. These advancements validate the efficacy of Mamba in visual tasks, leading to its application in image classification (Liu et al., 2024c; Zhu et al., 2024; Patro & Agneeswaran, 2024), video understanding (Wang et al., 2023b; Chen et al., 2024b; Li et al., 2024), image restoration (Guo et al., 2024; Shi et al., 2024; Zhen et al., 2024), point cloud analysis (Liang et al., 2024; Zhang et al., 2024; Liu et al., 2024b), and biomedical image segmentation (Ma et al., 2024; Xing et al., 2024; Liu et al., 2024a). These studies have demonstrated the effectiveness of state space models in visual tasks, providing a solid foundation for further exploration of their potential in visual perception.

3. Method

3.1. Exploring Mamba for Perception

3.1.1. Overall Framework

The overall architecture of our proposed QMamba model is depicted in Figure 2. To address the unique challenges of visual perception tasks, we developed a novel architecture that rethinks the design principles of state space models for visual data. The network incorporates a hierarchical residual structure, where convolutional layers enable effective feature extraction while specialized activation layers compute adaptive gating signals. Our architecture organizes the processing into multiple network stages, each combining a strategic downsampling layer with our enhanced Mamba-based processing block. This design enables the construction of multi-level representations at varying resolutions, facilitating the extraction of richer perceptual features through progressive abstraction.

To systematically investigate the relationship between model capacity and quality perception performance, we developed three distinct variants of our architecture: QMamba-Tiny, QMamba-Small, and QMamba-Base. Each variant maintains the core architectural principles while scaling in complexity, allowing us to explore the trade-offs between computational efficiency and perceptual accuracy across different application scenarios.

3.1.2. Perception with Local Scanning

While the original Mamba model excels in natural language processing tasks with inherently causal inputs, it faces challenges when applied to visual tasks due to the absence of spatial causality. In particular, it struggles to capture complex spatial dependencies among image pixels, which hinders its ability to model local distortions effectively.

To address this, VMamba (Liu et al., 2024c) introduces a bidirectional horizontal and vertical scanning strategy to convert 2D images into 1D sequences suitable for sequential modeling. Although this enables global pixel-level modeling, it disrupts the continuity of locally adjacent tokens, weakening the model's ability to perceive fine-grained distortions—an essential factor for image quality assessment (IQA).

Inspired by LocalMamba (Huang et al., 2024), we adopt a window-based scanning approach that performs horizon-

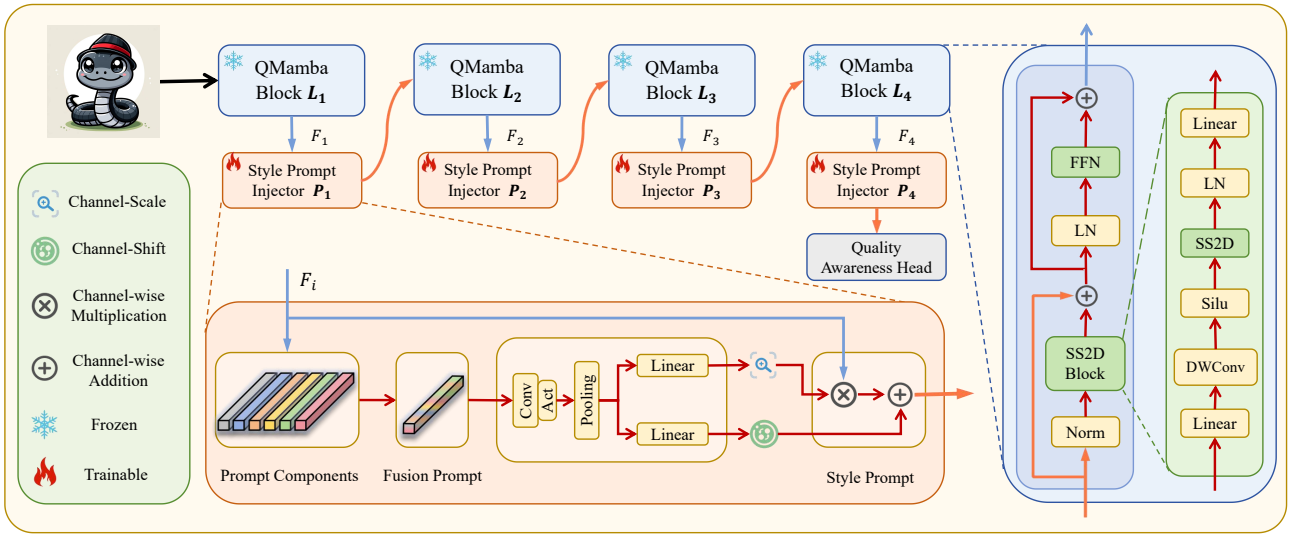

Figure 2. Architectural Overview of QMamba Framework and the Detailed of StylePrompt Tuning Mechanism.

tal scans within local windows, followed by window-level scans, and applies the same strategy vertically. This method enhances local distortion perception while retaining global awareness through hierarchical composition. Despite downsampling and aggregation in the input, this scanning scheme maintains a balance between local detail and broader context.

Unlike LocalMamba, which relies on attention-based dynamic routing and suffers from unstable inference and high computational cost, LQMamba adopts a hierarchical architecture with fixed-size windows that change with network depth. As illustrated in Figure 1 (b), this structure enables the model to capture multi-scale perceptual cues, from fine-grained distortions to broader contextual patterns, while maintaining stable and efficient inference. It achieves a good balance between accuracy and efficiency, making it well-suited for IQA tasks that demand consistent perception across various distortion types and spatial scales.

3.1.3. Analysis of Perceptual Capability in State Space Models

To investigate the feature selection mechanism of State Space Models (*i.e.*,Mamba) in visual quality assessment tasks, we employ t-SNE visualization for deep-layer feature analysis, as illustrated in Figure 3. Visualization results reveal that Mamba exhibits distinctive characteristic evolution patterns through dynamic state updating mechanisms, progressively enhancing distortion semantic perception across network hierarchies. Specifically, shallow layers preserve diverse features of original visual signals, while deeper layers adaptively filter redundant background information through gated state selection, intensifying distortion-type-specific feature focusing. This drives distortion-homogeneous samples to form cluster structures in feature space while amplifying inter-class discriminability. Such feature refinement strategy not only retains critical discriminative information for quality assessment but also significantly improves model sensitivity to quality degradation cues, thereby strengthening prediction robustness.

3.2. Tuning the Mamba with StylePrompt

Although the Mamba architecture reduces computational complexity compared to other models, achieving higher performance still requires a substantial number of parameters, which poses a challenge for efficient transferable learning, crucial for IQA tasks. We observe that domain shifts in IQA tasks tend to correlate with their feature statistics or style (Lu et al., 2022), such as mean and variance. Building on this insight, we propose a lightweight tuning strategy, StylePrompt, designed to adjust the mean and variance of the pre-trained QMamba features, thereby aligning them with the distortions and content types of the target domain. This approach enables us to achieve results comparable to full-parameter fine-tuning while using a minimal number of parameters, significantly enhancing both the efficiency and performance of QMamba in transferable IQA tasks. Figure 2 illustrates the StylePrompt, which consists of two components that will be described in detail below:

3.2.1. StylePrompt Generation (SPG)

We designed the StylePrompt Generation phase to facilitate the creation of prompts and their interaction with the original features. In a multi-stage network architecture, as images progress through each stage of the network, we learn a set of prompts $P_s \in \mathbb{R}^{N\times1\times1\times C}$, containing N prompt com-

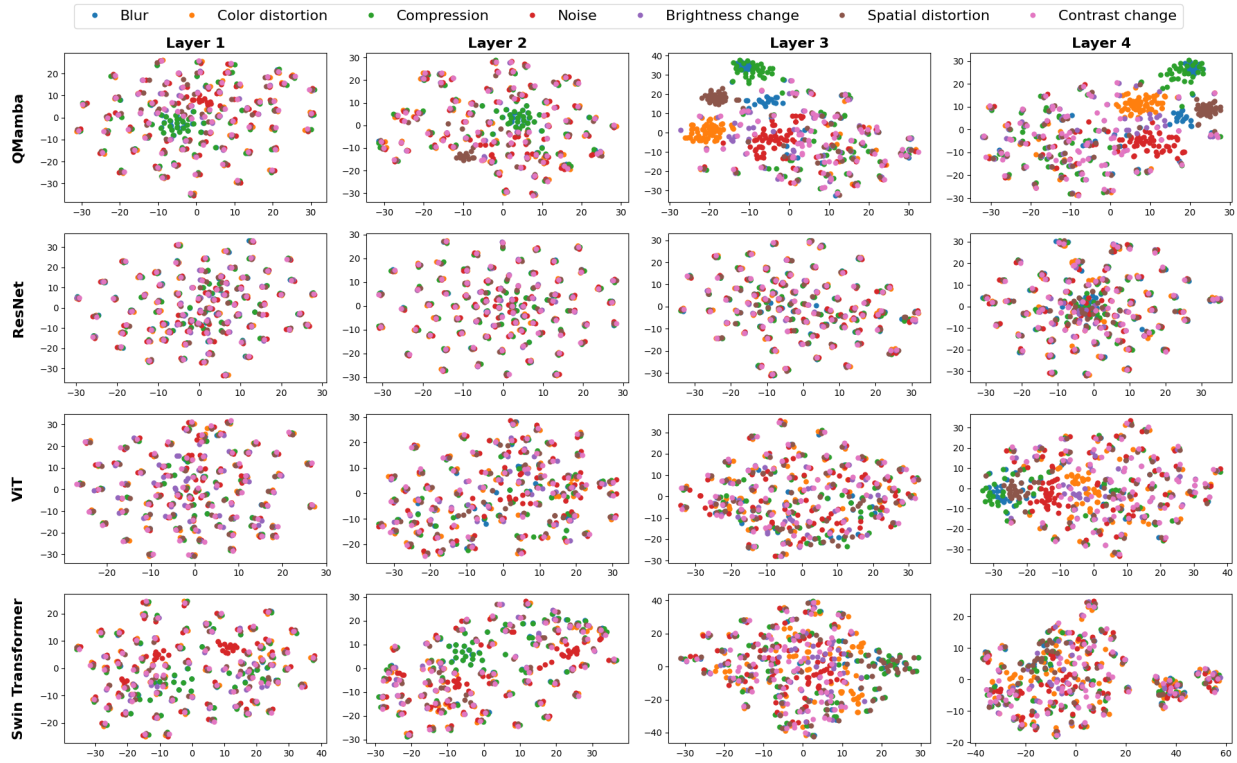

Figure 3. t-SNE Visualization of Distortion-Specific Feature Separation: QMamba vs. Conventional Backbones

ponents designed to generate affine parameters for style adaptation. These prompts are specifically utilized to inject the distortion information of the current data into the features $F_i \in \mathbb{R}^{\hat{H}\times\hat{W}\times\hat{C}}$, facilitating learning of the style pertinent to the target domain.

To enable the prompt components to extract specific style information from the input stream dynamically, we predict the weights for different prompt components based on the input features. This process involves performing a global pooling on the current layer's features followed by applying a softmax function to obtain the weights for the prompt group. These weights are then applied to the multiple prompt components to amalgamate them into a new prompt P_f, effectively encapsulating the current style information. The operation can be briefly summarized by the following formula:

$$P_f = \sum_{c=1}^{N} w_s P_s, \quad w_s = \text{Softmax}(\text{Conv1x1}(\text{GAP}(F_i))) \tag{1}$$

3.2.2. StylePrompt Injection (SPI)

In the SPG phase, the fused prompt P_f is created, containing the distortion style information of the target domain. During the subsequent StylePrompt Injection process, this style information is injected into the original features $F_i \in \mathbb{R}^{\hat{H}\times\hat{W}\times\hat{C}}$ produced by the current layer. To achieve this, P_f is matched to the channel dimensions of the current features using linear layers designed for dimensional alignment. The processed prompt then generates affine parameters $\gamma_v \in \mathbb{R}^{1\times1\times\hat{C}}$ and $\beta_v \in \mathbb{R}^{1\times1\times\hat{C}}$, which modulate the mean and variance of the original feature distribution solely along the channel dimension. This adaptive adjustment ensures that the feature distribution is tailored to the distortion style of the target domain with minimal computational overhead.

The process of our StylePrompt Injection can be summarized as follows:

$$\gamma_v = \text{Linear}_\gamma(\text{Conv}(P_f)) \tag{2}$$

$$\beta_v = \text{Linear}_\beta(\text{Conv}(P_f)) \tag{3}$$

$$F'_i = F_i \cdot (1 + \gamma_v) + \beta_v \tag{4}$$

Specifically, F'_i represents the original features after the injection of style information. These enhanced features will serve as the new input to the subsequent layer of the network.

		LIVE		CSIQ		TID2013		KADID		LIVEC		KonIQ		LIVEFB		SPAQ		
Method	GFLOPS	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	Average
ILNIQE	-	0.906	0.902	0.865	0.822	0.648	0.521	0.558	0.534	0.508	0.508	0.537	0.523	0.332	0.294	0.712	0.713	0.618
BRISQUE	-	0.944	0.929	0.748	0.812	0.571	0.626	0.567	0.528	0.629	0.629	0.685	0.681	0.341	0.303	0.817	0.809	0.664
WaDIQaM(5.24M)	2.43G	0.955	0.960	0.844	0.852	0.855	0.835	0.752	0.739	0.671	0.682	0.807	0.804	0.467	0.455	-	-	0.763
DBCNN(15.31M)	16.51G	0.971	0.968	**0.959**	**0.946**	0.865	0.816	0.856	0.851	0.869	0.851	0.884	0.875	0.551	0.545	0.915	0.911	0.852
TIQA(23.68M)	-	0.965	0.949	0.838	0.825	0.858	0.846	0.855	0.850	0.861	0.845	0.903	0.892	0.581	0.541	-	-	0.829
MetaIQA(13.24M)	1.82G	0.959	0.960	0.908	0.899	0.868	0.856	0.775	0.762	0.802	0.835	0.856	0.887	0.507	0.540	-	-	0.815
HyperIQA(27.38M)	4.31G	0.966	0.962	0.942	0.923	0.858	0.840	0.845	0.852	0.882	0.859	0.917	0.906	0.602	0.544	0.915	0.911	0.858
TReS(152.45M)	20.03G	0.968	0.969	0.942	0.922	0.883	0.863	0.858	0.859	0.877	0.846	0.928	0.915	0.625	0.554	-	-	0.858
MUSIQ(27.13M)	9.02G	0.911	0.940	0.893	0.871	0.815	0.773	0.872	0.875	0.746	0.702	0.928	0.916	0.661	0.566	0.921	0.918	0.832
DEIQT(24.04M)	5.41G	**0.982**	**0.980**	**0.963**	**0.946**	0.908	0.892	0.887	0.889	0.894	0.875	0.934	0.921	0.663	0.571	0.923	0.919	0.884
LoDa(*)	23.74G	**0.979**	**0.975**	-	-	0.901	0.869	0.936	0.931	0.899	0.876	0.944	0.932	**0.679**	**0.578**	0.928	0.925	0.882
ResNet-50(23.51M)	4.11G	0.879	0.884	0.861	0.841	0.747	0.686	0.784	0.786	0.868	0.831	0.908	0.886	0.313	0.269	0.907	0.907	0.772
ResNet-101(42.50M)	7.83G	0.918	0.921	0.891	0.867	0.779	0.727	0.722	0.719	0.862	0.824	0.918	0.904	0.420	0.347	0.908	0.906	0.790
ResNet-152(58.15M)	11.53G	0.926	0.927	0.923	0.899	0.765	0.717	0.764	0.760	0.859	0.816	0.919	0.898	0.433	0.353	0.907	0.907	0.798
ViT-T(5.52M)	1.26G	0.786	0.792	0.725	0.717	0.728	0.699	0.832	0.836	0.777	0.730	0.852	0.852	0.521	0.461	0.896	0.896	0.756
ViT-S(21.67M)	4.61G	0.900	0.896	0.832	0.815	0.873	0.859	0.893	0.894	0.831	0.799	0.922	0.905	0.539	0.443	0.919	0.917	0.827
ViT-B(85.80M)	17.58G	0.961	0.955	0.924	0.912	0.904	0.905	0.910	0.908	0.875	0.837	0.913	0.895	0.491	0.452	0.914	0.912	0.854
Swin-T(27.52M)	4.51G	0.879	0.883	0.865	0.847	0.937	0.925	0.923	0.922	0.880	0.845	0.901	0.881	0.476	0.453	0.922	0.919	0.841
Swin-S(48.84M)	8.77G	0.883	0.896	0.884	0.874	0.931	0.918	0.895	0.894	0.907	**0.884**	0.931	0.914	0.476	0.433	0.918	0.915	0.847
Swin-B(86.74M)	15.47G	0.945	0.948	0.941	0.935	0.942	0.933	0.934	0.932	0.892	0.858	0.945	0.932	0.507	0.471	0.923	0.921	0.872
QMamba-T (27.99M)	4.47G	0.959	0.959	0.940	0.918	0.951	0.945	0.934	0.930	0.898	0.866	0.941	0.925	0.675	0.581	**0.934**	**0.929**	0.893
QMamba-S (49.37M)	8.71G	0.962	0.965	0.921	0.903	**0.957**	**0.955**	0.934	**0.933**	0.903	0.874	0.943	0.930	**0.677**	**0.573**	0.932	0.927	0.893
QMamba-B (87.53M)	15.35G	0.960	0.961	0.908	0.889	0.953	0.949	0.935	0.932	**0.908**	0.876	0.943	0.930	0.675	0.579	0.933	0.929	0.891
LQMamba-T(29.87M)	4.44G	0.958	0.959	0.935	0.916	0.952	0.950	0.938	0.923	0.903	0.863	0.943	0.928	0.672	0.574	0.933	0.927	0.892
LQMamba-S(52.91M)	8.66G	0.962	0.964	0.933	0.914	0.955	0.949	**0.941**	0.928	0.907	0.882	**0.946**	**0.934**	0.676	0.574	0.933	0.929	**0.895**
LQMamba-B(93.79M)	15.30G	0.959	0.951	0.915	0.889	**0.965**	**0.964**	**0.943**	**0.941**	**0.913**	**0.888**	**0.947**	**0.933**	0.675	0.582	**0.934**	**0.929**	**0.896**

* LoDa has a total of 118.23M model parameters and 8.93M trainable parameters.

Table 1. Performance Comparison for Task-Specific IQA. Bold Indicates the Top Two Results.

4. Experiments

4.1. Experimental Setup

4.1.1. Datasets

We conducted foundational experiments on ten popular IQA datasets, which include four synthetic datasets: LIVE(Sheikh et al., 2006), CSIQ(Larson & Chandler, 2010), TID2013(Ponomarenko et al., 2015), and KADID(Lin et al., 2019); four authentic datasets: LIVEC(Ghadiyaram & Bovik, 2015), KonIQ(Hosu et al., 2020), LIVEFB(Ying et al., 2020), and SPAQ(Fang et al., 2020); and two AIGC datasets: AIGC2023(Wang et al., 2023a) and AGIQA3K(Li et al., 2023a). We will provide a more detailed introduction to these datasets in Appendix.

4.1.2. Evaluation Criteria

The evaluation metrics employed in our study are the widely utilized Pearson Linear Correlation Coefficient (PLCC) and Spearman's Rank Correlation Coefficient (SRCC), both of which range from 0 to 1. Values approaching 1 denote a higher degree of prediction relevance.

4.1.3. Experimental Details

Our experimental methodology closely follows the training strategy outlined in DEIQT(Qin et al., 2023), where input images are randomly cropped into ten patches, each with a resolution of 224×224. We employed three variants of the VMamba architecture: QMamba-B, QMamba-S, and QMamba-T. Both QMamba-B and QMamba-S feature an encoder with a depth of 15 blocks, with QMamba-B incorporating an embedding dimension of 128, and QMamba-S using an embedding dimension of 96. In contrast, QMamba-T is designed with a reduced depth of 4 blocks and an embedding dimension of 96. Training procedures leveraged weights pre-trained on the ImageNet-1K dataset, spanning a total of 9 epochs. Batch sizes were adjusted according to the respective dataset sizes, *e.g.*, 32 for LIVEC and 128 for KonIQ. We used the AdamW optimizer for training, with the learning rate set to 2×10^{-4} and a decay factor of 10 applied every 3 epochs. We compared the performance of ResNet(He et al., 2016), ViT(Dosovitskiy et al., 2020), and Swin Transformer(Liu et al., 2021), all of which were implemented using the official versions and loaded with

pre-trained weights. To ensure fairness, other experimental settings were kept as consistent as possible. All experiments were conducted using multiple NVIDIA RTX 4090 GPUs.

4.2. A Comparison Between Different IQA Backbones

4.2.1. TASK-SPECIFIC IQA

We conducted comprehensive training and testing across the ten datasets previously introduced, drawing analytical comparisons based on results reported by existing methods. For this task, 80% of the images in each dataset were used for training, while the remaining 20% were reserved for testing. Given the predominance of BIQA methods targeting synthetic and authentic datasets, Table 1 presents a detailed comparative analysis of state-of-the-art (SOTA) methods and popular architectures such as ResNet, ViT, and Swin Transformer, in comparison to our QMamba architecture. This comparison elucidates performance discrepancies and computational complexities across different parameter configurations. The results in the table demonstrate that the LQ-Mamba configuration achieves optimal performance across six datasets. Compared to existing IQA methods with similar parameter counts, QMamba-T exhibits lower GFLOPS and superior performance, confirming the efficiency and reduced computational complexity of the Mamba architecture. Comparative results for the two AIGC datasets are documented in the appendix.

Mixed Training				
LIVE & KADID & LIVEC & KonIQ & AGIQA3K & AIGCIQA2023				
Method	Parameters	GFLOPS	PLCC_Average	SRCC_Average
ResNet-50	23.51M	4.11G	0.878	0.853
ViT-S	21.67M	4.61G	0.891	0.867
Swin-T	27.52M	4.51G	0.900	0.883
DEIQT	24.04M	5.41G	0.895	0.873
LoDa	8.93M*	23.68G	0.876	0.855
QMamba-T	27.99M	4.47G	0.905	0.886
LQMamba-T	29.87M	4.44G	**0.909**	**0.888**

*Trainable parameters.

Table 2. Performance Comparison for Universal IQA.

4.2.2. UNIVERSAL IQA

We evaluated the effectiveness of the QMamba architecture for universal tasks by employing mixed training across six different datasets: two synthetic datasets (LIVE, KADID), two authentic datasets (LIVEC, KonIQ), and two AIGC datasets (AIGC2023, AGIQA3k). For each dataset, 20% of the data was reserved for performance testing. The average results of models with similar scales are presented in Table 2, with detailed results provided in Appendix. These findings highlight strong multi-tasking capabilities of Mamba. Compared to several mainstream models, QMamba performed well across most datasets, confirming its effectiveness in handling general tasks.

4.2.3. ANALYSIS

Our investigation into the efficacy of QMamba for IQA tasks reveals two key insights through analysis of the KADID dataset (7 distortion types) and cross-dataset validation. As shown in Figure 3, t-SNE visualization demonstrates superior distortion discrimination of QMamba: 1) Tightly clustered features for each distortion type, 2) Clear separation between dissimilar artifacts, and 3) Minimal inter-class overlap compared to ViT's partial merging and CNN/Swin architectures' significant feature entanglement.

This discriminative capability directly impacts practical performance. While QMamba demonstrates modest gains on simpler datasets such as LIVE and CSIQ, which contain only 4 to 5 distortion types, it achieves substantial improvements on more complex benchmarks like TID2013 and KADID, which include 24 to 25 distortion types. The architecture based on state-space modeling enables adaptive frequency processing through a selective scanning mechanism, dynamically emphasizing distortion-critical patterns while suppressing irrelevant features. This approach stands in contrast to convolutional networks with fixed receptive fields and Transformers that tend to over-mix local characteristics through global attention.

Our findings indicate that QMamba is well-suited for new quality assessment scenarios involving complex and mixed distortions, such as those introduced by neural compression or generative models. In these cases, traditional architectures often fail to capture subtle or entangled artifacts, while the selective modeling in QMamba provides better adaptability and robustness.

4.3. Efficient Transfer Learning for Mamba-Based IQA

In the context of transferable IQA tasks, we conducted domain-specific training using synthetic datasets (LIVE, KADID), authentic datasets (LIVEC, KonIQ), and AIGC datasets (AIGC2023, AGIQA3K). After training in one domain, models were directly transferred and tested on datasets from the other two domains. We employed the StylePrompt technique, as illustrated in Figure 2, where the architecture was kept intact by freezing all model parameters and fine-tuning only the StylePrompt module, which involved approximately 4% of the total parameter count. This approach achieved performance levels comparable to those obtained through full-parameter training. The outcomes, presented in Table 3, clearly demonstrate the effectiveness and efficiency of the proposed StylePrompt method for transferable IQA tasks. The value of "Average" represents the mean performance across all domain transfers, with more detailed results provided in the Appendix.

Train	KonIQ & LIVEC (Authentic)								KADID & LIVE (Synthetic)								
Test	KADID		LIVE		AIGC2023		AGIQA3K		KonIQ		LIVEC		AIGC2023		AGIQA3K		
Fine-tuning Method	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	Average
DEIQT	0.583	0.595	0.558	0.564	0.802	0.794	0.698	0.632	0.518	0.526	0.579	0.532	0.580	0.575	0.512	0.506	0.601
LoDa	0.600	0.600	0.733	0.729	0.808	0.800	0.715	0.650	0.527	0.513	0.558	0.499	0.592	0.586	0.564	0.541	0.622
Without_tuning	0.535	0.501	0.726	0.744	0.789	0.784	0.726	0.744	0.553	0.549	0.595	0.550	0.600	0.612	0.675	0.649	0.642
Lin_Probe(R)	0.606	0.581	0.783	0.811	0.824	0.804	0.769	0.695	0.871	0.848	0.785	0.751	0.803	0.788	0.771	0.723	0.755
Full_tuning **(93.79M)**	0.936	0.930	0.941	0.940	0.885	0.863	0.910	0.851	0.943	0.928	0.899	0.873	0.880	0.859	0.910	0.856	**0.908**
StylePrompt **(3.83M)**	0.920	0.912	0.949	0.948	0.877	0.854	0.908	0.854	0.932	0.913	0.888	0.866	0.880	0.865	0.906	0.852	**0.901**
StylePrompt & R	0.921	0.913	0.944	0.945	0.877	0.853	0.906	0.847	0.931	0.912	0.889	0.860	0.876	0.857	0.905	0.849	0.898

Table 3. Performance Comparison for Transferable IQA.

4.4. Ablation Study

4.4.1. Different Scanning Method

As shown in the task-specific results (Table 1) and universal evaluation results (Table 2), models adopting local scanning consistently outperform those based on cross-scanning strategies. This motivates our exploration of local scanning mechanisms, as employed in LQMamba.

Although the average performance gap between QMamba and LQMamba may appear small, further analysis reveals that LQMamba performs better on most individual datasets. The marginal overall gain is primarily due to relatively lower improvements on simpler datasets such as LIVE and CSIQ, which contain fewer distortion types and less diverse content. In contrast, on more challenging datasets like TID2013 and KADID-10k, which feature a wide variety of fine-grained distortions, LQMamba shows clear advantages (*e.g.*, SRCC on TID2013: 0.964 vs. 0.949; on KADID: 0.941 vs. 0.932). These results highlight the effectiveness of the local scanning design in complex, distortion-rich scenarios where precise modeling of local artifacts is critical.

4.4.2. Different Model Scale

In our empirical analysis, as documented in Table 1, we discern that while QMamba-Base exhibits exceptionally robust quality perception capabilities, QMamba-Small either matches or exceeds the performance of QMamba-Base across the majority of the datasets. Although QMamba-Tiny displays a modest decline in performance metrics, it still delivers results that are competitive with current SOTA methods, solely utilizing the capabilities of QMamba-Tiny.

Tuning Strategy	Parameters	PLCC_Avg.	SRCC_Avg.
SSF	6.1M	0.750	0.735
Crossattn_Prompt	12.17M	0.806	0.772
Conv_Prompt	28.33M	0.883	0.856
StylePrompt**(ours)**	3.83M	**0.911**	**0.890**

Table 4. Ablation Study on Different Prompt Tuning Strategies

4.4.3. Different Prompt Tuning Strategies

To validate the effectiveness of the StylePrompt Generation (SPG) process, we conducted tests by directly learning a set of affine parameters γ and β to modulate the original features, rather than using prompts for learning, similar to the SSF(Lian et al., 2022). Additionally, to assess the effectiveness of the StylePrompt interaction method, we explored various interaction strategies during tuning, including convolutional prompt interaction and cross-attention prompt interaction. The results in Table 4 demonstrate the superior efficiency and performance of StylePrompt.

In addition, we conducted several ablation studies related to the design of the prompts, including variations in the number and shape of the prompts. The detailed results of these experiments are provided in Appendix.

4.5. Discussion and Future Work

Looking ahead, the sequential modeling nature and efficiency of our SSM-based architecture make it a promising candidate for extension to video and audio quality assessment. Given the higher temporal complexity in VQA and the inherent sequential structure of audio signals, our method offers a unified and lightweight foundation for future research across visual and multimodal quality evaluation tasks.

5. Conclusion

In this paper, we introduced QMamba, a novel state space framework for image quality assessment that integrates task-specific evaluation, universal perception, and cross-domain transferability. Extensive experiments demonstrate that QMamba consistently outperforms established vision models, achieving significant improvements in accuracy while requiring substantially lower computational resources. The proposed StylePrompt mechanism enables robust cross-domain adaptation through lightweight dynamic feature recalibration, setting new benchmarks for IQA and demonstrating strong practical potential in real-world applications.

Acknowledgement

This work was supported in part by NSFC under Grant 623B2098, 62021001, and 62371434, as well as the China Postdoctoral Science Foundation-Anhui Joint Support Program under Grant Number 2024T017AH.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Chen, C., Mo, J., Hou, J., Wu, H., Liao, L., Sun, W., Yan, Q., and Lin, W. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 2024a.

Chen, G., Huang, Y., Xu, J., Pei, B., Chen, Z., Li, Z., Wang, J., Li, K., Lu, T., and Wang, L. Video mamba suite: State space model as a versatile alternative for video understanding. *arXiv preprint arXiv:2403.09626*, 2024b.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Fang, Y., Zhu, H., Zeng, Y., Ma, K., and Wang, Z. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3677–3686, 2020.

Fei, B., Lyu, Z., Pan, L., Zhang, J., Yang, W., Luo, T., Zhang, B., and Dai, B. Generative diffusion prior for unified image restoration and enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9935–9946, 2023.

Fu, D. Y., Dao, T., Saab, K. K., Thomas, A. W., Rudra, A., and Ré, C. Hungry hungry hippos: Towards language modeling with state space models. *arXiv preprint arXiv:2212.14052*, 2022.

Ghadiyaram, D. and Bovik, A. C. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, 2015.

Golestaneh, S. A., Dadsetan, S., and Kitani, K. M. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1220–1230, 2022.

Gu, A. and Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Gu, A., Dao, T., Ermon, S., Rudra, A., and Ré, C. Hippo: Recurrent memory with optimal polynomial projections. *Advances in neural information processing systems*, 33: 1474–1487, 2020.

Gu, A., Goel, K., and Ré, C. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

Guo, H., Li, J., Dai, T., Ouyang, Z., Ren, X., and Xia, S.-T. Mambair: A simple baseline for image restoration with state-space model. *arXiv preprint arXiv:2402.15648*, 2024.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hosu, V., Lin, H., Sziranyi, T., and Saupe, D. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020.

Huang, T., Pei, X., You, S., Wang, F., Qian, C., and Xu, C. Localmamba: Visual state space model with windowed selective scan. *arXiv preprint arXiv:2403.09338*, 2024.

Kang, L., Ye, P., Li, Y., and Doermann, D. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1733–1740, 2014.

Ke, J., Wang, Q., Wang, Y., Milanfar, P., and Yang, F. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5148–5157, 2021.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Larson, E. C. and Chandler, D. M. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of electronic imaging*, 19(1):011006–011006, 2010.

Li, C., Zhang, Z., Wu, H., Sun, W., Min, X., Liu, X., Zhai, G., and Lin, W. Agiqa-3k: An open database for ai-generated image quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023a.

Li, K., Li, X., Wang, Y., He, Y., Wang, Y., Wang, L., and Qiao, Y. Videomamba: State space model for efficient video understanding. *arXiv preprint arXiv:2403.06977*, 2024.

Li, X., Shi, J., and Chen, Z. Task-driven semantic coding via reinforcement learning. *IEEE Transactions on Image Processing*, 30:6307–6320, 2021.

Li, X., Li, B., Jin, X., Lan, C., and Chen, Z. Learning distortion invariant representation for image restoration from a causality perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1714–1724, 2023b.

Li, X., Lu, Y., and Chen, Z. Freqalign: Excavating perception-oriented transferability for blind image quality assessment from a frequency perspective. *IEEE Transactions on Multimedia*, 2023c.

Lian, D., Zhou, D., Feng, J., and Wang, X. Scaling & shifting your features: A new baseline for efficient model tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Liang, D., Zhou, X., Wang, X., Zhu, X., Xu, W., Zou, Z., Ye, X., and Bai, X. Pointmamba: A simple state space model for point cloud analysis. *arXiv preprint arXiv:2402.10739*, 2024.

Lin, H., Hosu, V., and Saupe, D. Kadid-10k: A large-scale artificially distorted iqa database. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–3. IEEE, 2019.

Liu, J., Zhou, W., Li, X., Xu, J., and Chen, Z. Liqa: Lifelong blind image quality assessment. *IEEE Transactions on Multimedia*, 2022.

Liu, J., Yang, H., Zhou, H.-Y., Xi, Y., Yu, L., Yu, Y., Liang, Y., Shi, G., Zhang, S., Zheng, H., et al. Swin-umamba: Mamba-based unet with imagenet-based pretraining. *arXiv preprint arXiv:2402.03302*, 2024a.

Liu, J., Yu, R., Wang, Y., Zheng, Y., Deng, T., Ye, W., and Wang, H. Point mamba: A novel point cloud backbone based on state space model with octree-based ordering strategy. *arXiv preprint arXiv:2403.06467*, 2024b.

Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., and Liu, Y. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024c.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Lu, Y., Li, X., Liu, J., and Chen, Z. Styleam: Perception-oriented unsupervised domain adaption for non-reference image quality assessment. *arXiv preprint arXiv:2207.14489*, 2022.

Lu, Y., Li, X., Pei, Y., Yuan, K., Xie, Q., Qu, Y., Sun, M., Zhou, C., and Chen, Z. Kvq: Kwai video quality assessment for short-form videos. *CVPR*, 2024.

Ma, J., Li, F., and Wang, B. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024.

Min, X., Zhai, G., Gu, K., Liu, Y., and Yang, X. Blind image quality estimation via distortion aggravation. *IEEE Transactions on Broadcasting*, 64(2):508–517, 2018.

Mittal, A., Moorthy, A. K., and Bovik, A. C. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012a.

Mittal, A., Soundararajan, R., and Bovik, A. C. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012b.

Patro, B. N. and Agneeswaran, V. S. Simba: Simplified mamba-based architecture for vision and multivariate time series. *arXiv preprint arXiv:2403.15360*, 2024.

Ponomarenko, N., Jin, L., Ieremeiev, O., Lukin, V., Egiazarian, K., Astola, J., Vozel, B., Chehdi, K., Carli, M., Battisti, F., et al. Image database tid2013: Peculiarities, results and perspectives. *Signal processing: Image communication*, 30:57–77, 2015.

Qin, G., Hu, R., Liu, Y., Zheng, X., Liu, H., Li, X., and Zhang, Y. Data-efficient image quality assessment with attention-panel decoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2091–2100, 2023.

Saad, M. A., Bovik, A. C., and Charrier, C. Blind image quality assessment: A natural scene statistics approach in the dct domain. *IEEE transactions on Image Processing*, 21(8):3339–3352, 2012.

Saha, A., Mishra, S., and Bovik, A. C. Re-iqa: Unsupervised learning for image quality assessment in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5846–5855, 2023.

Sheikh, H. R., Sabir, M. F., and Bovik, A. C. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing*, 15(11):3440–3451, 2006.

Shi, Y., Xia, B., Jin, X., Wang, X., Zhao, T., Xia, X., Xiao, X., and Yang, W. Vmambair: Visual state space model for image restoration. *arXiv preprint arXiv:2403.11423*, 2024.

Shin, N.-H., Lee, S.-H., and Kim, C.-S. Blind image quality assessment based on geometric order learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12799–12808, 2024.

Smith, J. T., Warrington, A., and Linderman, S. W. Simplified state space layers for sequence modeling. *arXiv preprint arXiv:2208.04933*, 2022.

Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., and Zhang, Y. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3667–3676, 2020.

Talebi, H. and Milanfar, P. Nima: Neural image assessment. *IEEE transactions on image processing*, 27(8): 3998–4011, 2018.

Tu, Z., Wang, Y., Birkbeck, N., Adsumilli, B., and Bovik, A. C. Ugc-vqa: Benchmarking blind video quality assessment for user generated content. *IEEE Transactions on Image Processing*, 30:4449–4464, 2021.

Venkatanath, N., Praneeth, D., Bh, M. C., Channappayya, S. S., and Medasani, S. S. Blind image quality evaluation using perception based features. In *2015 twenty first national conference on communications (NCC)*, pp. 1–6. IEEE, 2015.

Wang, J., Duan, H., Liu, J., Chen, S., Min, X., and Zhai, G. Aigciqa2023: A large-scale image quality assessment database for ai generated images: from the perspectives of quality, authenticity and correspondence. In *CAAI International Conference on Artificial Intelligence*, pp. 46–57. Springer, 2023a.

Wang, J., Zhu, W., Wang, P., Yu, X., Liu, L., Omar, M., and Hamid, R. Selective structured state-spaces for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6387–6397, 2023b.

Wu, Y., Li, X., Zhang, Z., Jin, X., and Chen, Z. Learned block-based hybrid image compression. *IEEE Transactions on Circuits and Systems for Video Technology*, 32 (6):3978–3990, 2021.

Xing, Z., Ye, T., Yang, Y., Liu, G., and Zhu, L. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. *arXiv preprint arXiv:2401.13560*, 2024.

Ying, Z., Niu, H., Gupta, P., Mahajan, D., Ghadiyaram, D., and Bovik, A. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3575–3585, 2020.

Yu, Z., Guan, F., Lu, Y., Li, X., and Chen, Z. Sf-iqa: Quality and similarity integration for ai generated image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6692–6701, 2024a.

Yu, Z., Guan, F., Lu, Y., Li, X., and Chen, Z. Video quality assessment based on swin transformerv2 and coarse to fine strategy. *arXiv preprint arXiv:2401.08522*, 2024b.

Zeng, H., Zhang, L., and Bovik, A. C. A probabilistic quality representation approach to deep blind image quality prediction. *arXiv preprint arXiv:1708.08190*, 2017.

Zhang, T., Li, X., Yuan, H., Ji, S., and Yan, S. Point could mamba: Point cloud learning via state space model. *arXiv preprint arXiv:2403.00762*, 2024.

Zhang, W., Ma, K., Yan, J., Deng, D., and Wang, Z. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1):36–47, 2020.

Zhang, W., Zhai, G., Wei, Y., Yang, X., and Ma, K. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14071–14081, 2023.

Zhao, K., Yuan, K., Sun, M., Li, M., and Wen, X. Quality-aware pre-trained models for blind image quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22302–22313, 2023.

Zhen, Z., Hu, Y., and Feng, Z. Freqmamba: Viewing mamba from a frequency perspective for image deraining. *arXiv preprint arXiv:2404.09476*, 2024.

Zhu, H., Li, L., Wu, J., Dong, W., and Shi, G. Metaiqa: Deep meta-learning for no-reference image quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14143–14152, 2020.

Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., and Wang, X. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.

A. Appendix / supplemental material

A.1. Detailed Description of the Datasets

We evaluate the performance of the proposed method across eight widely recognized BIQA datasets, comprising both synthetic and authentic datasets. The synthetic datasets include LIVE(Sheikh et al., 2006), CSIQ(Larson & Chandler, 2010), TID2013(Ponomarenko et al., 2015), and KADID-10k(Lin et al., 2019). These datasets feature a small number of pristine images that are synthetically distorted using various techniques such as JPEG compression and Gaussian blurring. Specifically, LIVE contains 799 images with 5 types of distortion, CSIQ includes 866 images with 6 distortion types, TID2013 comprises 3000 images with 24 distortion types, and KADID-10k includes 10125 images with 25 distortion types.

On the other hand, the authentic datasets include LIVEC(Ghadiyaram & Bovik, 2015), KonIQ-10k(Hosu et al., 2020), SPAQ(Fang et al., 2020), and FLIVE(Ying et al., 2020). LIVEC consists of 1,162 images with diverse authentic distortions captured by mobile devices. KonIQ-10k is composed of 10,073 images selected from YFCC-100M, covering a wide range of distortions such as brightness, colorfulness, contrast, noise, and sharpness. SPAQ contains 11,125 images collected by various smartphones, representing a large variety of scene categories. FLIVE, the largest in-the-wild IQA dataset to date, contains 39,810 real-world images with diverse content, sizes, and aspect ratios.

In response to the rapid development of AI-generated content, we also employed two additional datasets: AIGCIQA2023(Wang et al., 2023a) and AGIQA3K(Li et al., 2023a). AIGCIQA2023 contains over 2000 images generated by six state-of-the-art text-to-image models, evaluated through a subjective experiment on quality, authenticity, and correspondence. AGIQA3K consists of 2,982 images from GAN, autoregression, and diffusion-based models, with annotations for perceptual quality and text-to-image alignment.

A.2. Additional Experimental Results

A.2.1. Task-specific IQA results on AIGC datasets

We present the test results of various backbones on the AIGC datasets in Table 5, with bold indicating the best results.

A.2.2. Universal IQA detailed data

We provide the detailed results of Universal IQA in Table 6.

Model		AGIQA3K		AIGCIQA2023		
	GFLOPS	PLCC	SRCC	PLCC	SRCC	Average
ResNet-50(23.51M)	4.11G	0.901	0.840	0.795	0.797	0.833
ResNet-101(42.50M)	7.83G	0.907	0.847	0.834	0.831	0.855
ResNet-152(58.15M)	11.53G	0.901	0.832	0.841	0.834	0.852
ViT-T(5.52M)	1.26G	0.865	0.787	0.760	0.766	0.795
ViT-S(21.67M)	4.61G	0.891	0.819	0.842	0.822	0.844
ViT-B(85.80M)	17.58G	0.897	0.830	0.853	0.835	0.854
Swin-T(27.52M)	4.51G	0.906	0.847	0.867	0.844	0.866
Swin-S(48.84M)	8.77G	0.908	0.849	0.875	0.857	0.872
Swin-B(86.74M)	15.47G	0.909	0.852	0.886	0.863	0.878
QMamba-T (27.99M)	4.47G	0.913	0.858	0.888	0.873	0.883
QMamba-S (49.37M)	8.71G	0.912	0.858	**0.889**	**0.875**	**0.884**
QMamba-B (87.53M)	15.35G	0.914	0.861	0.886	0.868	0.882
LQMamba-T(29.87M)	4.44G	0.914	0.862	0.884	0.868	0.882
LQMamba-S(52.91M)	8.66G	0.913	**0.864**	0.888	0.869	**0.884**
LQMamba-B(93.79M)	15.30G	**0.915**	0.858	0.888	0.871	0.883

Table 5. Results of AIGC for task-specific IQA on various backbones

A.2.3. Transferable IQA results trained on the AIGC domain

We provide the detailed results of Transferable IQA trained in the AIGC domain in Table 7.

A.2.4. Detailed results of ablation studies

- **Ablation study on different prompt tuning strategies.**
  We have provided the detailed experimental data for the ablation study on different prompt tuning strategies in Table 8.

- **Ablation study results for different prompt designs.**
  We investigated the impact of the number of prompt components on performance and found that setting the number to six yields optimal results. Additionally, we explored whether varying the spatial dimensions and sizes of prompts would enhance performance. Our findings show that a spatial size of (1,1), focusing solely on the channel dimension, offers the best results. The outcomes of these ablation studies across multiple datasets are presented in Tables 9 and 10.

A.3. Limitations

Q-Mamba heavily relies on pre-trained weights from ImageNet-1K, which may limit its applicability to domains with significantly different data distributions. Future work could explore pre-training on more diverse datasets to improve generalization capabilities. Despite the efficiency improvements brought by the StylePrompt, the computa-

Train	LIVE & KADID & LIVEC & KonIQ & AGIQA3K & AIGCIQA2023														
Test		LIVE		KADID		LIVEC		KonIQ		AGIQA3K		AIGCIQA2023			
	GFLOPS	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	Average
ResNet-50	4.11G	0.916	0.908	0.868	0.869	0.885	0.841	0.896	0.869	0.884	0.813	0.821	0.815	0.865
ViT-S	4.61G	0.921	0.918	0.920	0.917	0.883	0.828	0.911	0.892	0.885	0.822	0.825	0.824	0.879
Swin-T	4.51G	0.927	**0.927**	0.925	0.921	0.889	0.863	0.929	0.918	0.897	0.840	0.834	0.828	0.892
DEIQT	4.68G	0.907	0.906	0.898	0.895	0.896	0.855	0.928	0.906	0.902	0.841	0.839	**0.835**	0.884
LoDa	23.68G	0.895	0.902	0.907	0.900	0.872	0.842	0.900	0.875	0.874	0.809	0.809	0.801	0.866
QMamba-T	4.47G	0.923	0.923	0.938	0.932	0.898	0.863	0.932	0.917	0.906	0.851	0.835	0.830	0.896
LQMamba-T	4.44G	**0.929**	0.926	**0.943**	**0.939**	**0.899**	**0.863**	**0.936**	**0.921**	**0.908**	**0.853**	**0.840**	0.828	**0.899**

Table 6. Performance comparison for universal IQA.

Train	AIGC2023 & AGIQA3K (AIGC)								
Test	KonIQ		LIVEC		KADID		LIVE		
Fine-tuning Method	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	Average
DEIQT	0.578	0.505	0.684	0.638	0.498	0.464	0.752	0.750	0.601
LoDa	0.564	0.500	0.542	0.548	0.555	0.524	0.840	0.833	0.622
Without_tuning	0.636	0.587	0.667	0.637	0.460	0.425	0.815	0.838	0.642
Lin_Probe(R)	0.818	0.792	0.790	0.761	0.570	0.540	0.799	0.826	0.755
Full_tuning **(93.79M)**	0.943	0.927	0.910	0.878	0.937	0.932	0.938	0.933	**0.908**
StylePrompt **(3.83M)**	0.929	0.909	0.902	0.871	0.918	0.913	0.926	0.922	**0.901**
StylePrompt & R	0.929	0.909	0.899	0.859	0.918	0.911	0.926	0.927	0.898

Table 7. Performance comparison for transferable IQA (trained on AIGC).

Train	KonIQ & LIVEC								KADID & LIVE								AIGC2023 & AGIQA3K								
Test	KADID		LIVE		AIGC2023		AGIQA3K		KonIQ		LIVEC		AIGC2023		AGIQA3K		KonIQ		LIVEC		KADID		LIVE		
Style and Prompt method	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	Average
SSF**(6.1M)**	0.718	0.700	0.735	0.769	0.804	0.801	0.714	0.667	0.863	0.829	0.643	0.598	0.668	0.675	0.714	0.685	0.866	0.828	0.730	0.706	0.724	0.708	0.820	0.856	0.743
Crossattn_Prompt**(12.17M)**	0.843	0.830	0.898	0.897	0.783	0.752	0.821	0.706	0.843	0.816	0.641	0.612	0.802	0.782	0.833	0.726	0.842	0.818	0.644	0.616	0.831	0.815	0.894	0.897	0.789
Conv_Prompt**(28.33M)**	0.911	0.910	0.945	0.946	**0.889**	**0.820**	0.890	0.825	0.897	0.881	0.797	0.761	0.864	0.827	0.861	0.828	0.898	0.884	0.805	0.756	0.901	0.899	**0.932**	**0.933**	0.869
StylePrompt**(ours)(3.83M)**	**0.920**	**0.912**	**0.949**	**0.948**	0.877	0.854	**0.908**	**0.854**	**0.932**	**0.913**	**0.888**	**0.866**	**0.880**	**0.865**	**0.906**	**0.852**	**0.929**	**0.909**	**0.902**	**0.871**	**0.918**	**0.913**	0.926	0.922	**0.901**

Table 8. Ablation study on different prompt tuning strategies

tional demands for training and deploying QMamba on very large-scale datasets or in real-time applications might still be substantial. Investigating methods to further reduce computational complexity without sacrificing performance could be beneficial. The current study focuses on specific types of distortions, and there may be other types of distortions that have not been sufficiently explored. Expanding the evaluation to cover a broader range of distortions could provide a more comprehensive validation of QMamba's robustness. Given that synthetic datasets might not fully capture the complexity and variability of authentic data, there is a risk of overfitting to these synthetic examples. Further evaluations on more diverse and extensive authentic datasets would help ensure the model's robustness and practical applicability. While QMamba shows promising results in cross-domain transferability, its effectiveness across vastly different domains (e.g., medical imaging versus natural images) has not been thoroughly validated. Further studies are needed to test and possibly adapt QMamba for such diverse applications.

Train	KonIQ & LIVEC								KADID & LIVE								AIGC2023 & AGIQA3K								
Test	KADID		LIVE		AIGC2023		AGIQA3K		KonIQ		LIVEC		AIGC2023		AGIQA3K		KonIQ		LIVEC		KADID		LIVE		
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	Average
N=1	0.914	0.907	0.943	0.944	0.877	0.850	0.900	0.833	0.924	0.906	0.873	0.846	0.872	0.850	0.898	0.834	0.924	0.907	0.890	0.855	0.904	0.899	**0.940**	**0.937**	0.8928
N=2	0.913	0.905	0.938	0.939	0.874	0.851	0.895	0.837	0.930	0.912	0.886	0.853	0.871	0.851	0.892	0.835	0.927	0.905	0.902	0.865	0.911	0.906	0.926	0.922	0.8936
N=4	0.920	0.913	0.944	0.949	0.875	0.851	0.904	0.840	0.929	0.913	0.877	0.849	0.876	0.857	0.904	0.842	0.927	0.904	0.885	0.847	0.913	0.907	0.925	0.931	0.8951
N=6	0.920	0.912	**0.949**	**0.948**	0.877	0.854	**0.908**	**0.854**	**0.932**	**0.913**	**0.888**	**0.866**	**0.880**	**0.865**	**0.906**	**0.852**	**0.929**	**0.909**	**0.902**	**0.871**	**0.918**	**0.913**	0.926	0.922	**0.9006**
N=8	**0.920**	**0.914**	0.930	0.934	0.874	0.851	0.903	0.841	0.930	0.912	0.873	0.843	0.876	0.854	0.899	0.841	0.927	0.906	0.889	0.859	0.918	0.912	0.922	0.926	0.8939
N=10	0.919	0.913	0.930	0.936	**0.877**	**0.856**	0.901	0.840	0.929	0.910	0.880	0.849	0.877	0.853	0.900	0.840	0.925	0.906	0.889	0.851	0.918	0.914	0.923	0.924	0.8942

Table 9. Ablation study results for different numbers of prompts (N).

Train	KonIQ & LIVEC								KADID & LIVE								AIGC2023 & AGIQA3K								
Test	KADID		LIVE		AIGC2023		AGIQA3K		KonIQ		LIVEC		AIGC2023		AGIQA3K		KonIQ		LIVEC		KADID		LIVE		
(H,W)	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	Average
(1,1)	0.920	0.912	**0.949**	**0.948**	0.877	0.854	**0.908**	**0.854**	**0.932**	**0.913**	**0.888**	**0.866**	**0.880**	**0.865**	**0.906**	**0.852**	**0.929**	**0.909**	**0.902**	**0.871**	0.918	0.913	0.926	0.922	**0.901**
(7,7)	**0.924**	**0.917**	0.947	0.948	0.876	0.852	0.906	0.844	0.931	0.912	0.881	0.848	0.875	0.857	0.902	0.843	0.928	0.906	0.890	0.855	**0.921**	**0.915**	0.927	0.931	0.897
(14,14)	**0.924**	**0.917**	0.947	0.948	0.876	0.852	0.906	0.844	0.931	0.912	0.881	0.848	0.875	0.857	0.902	0.843	0.928	0.906	0.890	0.855	**0.921**	**0.915**	0.927	0.931	0.897
(28,28)	0.917	0.909	0.948	0.947	0.871	0.849	0.905	0.847	0.929	0.909	0.876	0.845	0.873	0.855	0.903	0.841	0.928	0.905	0.895	0.860	0.902	0.898	**0.932**	**0.931**	0.895
layer_HW	0.921	0.914	0.948	0.948	**0.880**	**0.854**	0.905	0.844	0.929	0.909	0.876	0.850	0.879	0.859	0.903	0.844	0.928	0.906	0.891	0.858	0.913	0.908	0.920	0.919	0.896

Table 10. Ablation study for different prompt shape