PreResQ-R1: Response-Preference Disentangled Ranking-and-Scoring Reinforcement Optimization for Robust Visual Quality Assessment

Zehui Feng[1], Weichuan Wang[1], Xiaohan Chen[1], Ting Han[1,2,3,*]
[1]School of Design, Shanghai Jiao Tong University, Shanghai, China
[2]Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, China
[3]Innovation Center of Yangtze River Delta, Zhejiang University, Jiaxing, Zhejiang
{fzh_sjtu, weichuanwang, chen_11, hanting}@sjtu.edu.cn

Abstract

Visual Quality Assessment (QA) in multimodal large language models (MLLMs) jointly predicts absolute quality scores and relative perceptual rankings. Existing methods optimize score regression or ranking within a unified formulation, leading to unstable predictions and weak generalization under stochastic generation. We find this instability is amplified by response variability in MLLM reasoning, where stochastic decoding produces inconsistent score distributions and degrades cross-dataset robustness. This indicates QA depends not only on preference alignment, but also on response-level stability. We propose PreResQ-R1, a Response-Preference Disentangled Ranking-and-Scoring Reinforcement Optimization framework. It decouples response stabilization and preference learning, with a dual-branch reward modeling intra-sample consistency and inter-sample score-ranking alignment. For video QA, we introduce temporal-spatial evidence aggregation without dense motion modeling. PreResQ-R1 achieves state-of-the-art performance, outperforming comparable methods by 5.60% on IQA benchmarks and 2.53% on VQA benchmarks.

1 Introduction

Visual Quality Assessment (QA) aims to predict human perceptual judgments of visual fidelity without reference images (Wang et al., 2004), and is fundamental to applications such as image enhancement (Chen et al., 2024), image generation (Chen et al., 2025b), and computational photography (Cai et al., 2024). The absence of reference images makes perceptual inference challenging in robustness and interpretability. A key limitation lies in the joint optimization of absolute score regression and relative preference ranking within a single objective (Wu et al., 2025a). Regression-based methods provide calibrated scores but are sensitive to dataset-specific distributions, leading to poor cross-dataset generalization (Ke et al., 2021; You et al., 2025), while ranking-based methods improve relative ordering (Ying et al., 2019; Liu et al., 2017) but lack absolute supervision, resulting in suboptimal calibration.

Recent multimodal large language model (MLLM) based QA methods attempt to address this limitation by incorporating reasoning through chain-of-thought supervision and reinforcement learning (RL) (Touvron et al., 2023; Bai et al., 2025; Wang et al., 2025a). However, they still rely on unified scalar rewards without explicitly disentangling perceptual factors, and therefore fail to resolve the underlying objective mismatch. As a result, RL optimization suffers from noisy and unstable gradients across stochastic generations, leading to slow convergence and brittle generalization, even when advanced policy optimization schemes are employed (DeepSeek-AI, 2025).

Importantly, this instability is not merely an optimization artifact (Li et al., 2026d,c, 2024); it reflects an inherent property of QA under stochastic generation. Responses to the same input can vary substantially, directly altering the learned score distribution. Left unconstrained, such variability harms calibration and weakens cross-dataset generalization. By contrast, response-level rational consistency acts as an implicit regularizer, improving the stability and reliability of quality estimation. These findings suggest that beyond preference alignment, response stability is essential for distributional robustness under limited supervision.

Extending reasoning-oriented QA models to video further amplifies this misalignment between perceptual requirements and model design (Li et al., 2026a,b). Existing methods (Cao et al., 2025a) often rely on explicit motion modeling or dense temporal sampling, incurring additional computational cost and limited alignment with MLLM-based IQA reasoning. In contrast, human perception of video quality aggregates sparse spatial and temporal cues

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 7998–8024
October 24-29, 2026 ©2026 Association for Computational Linguistics

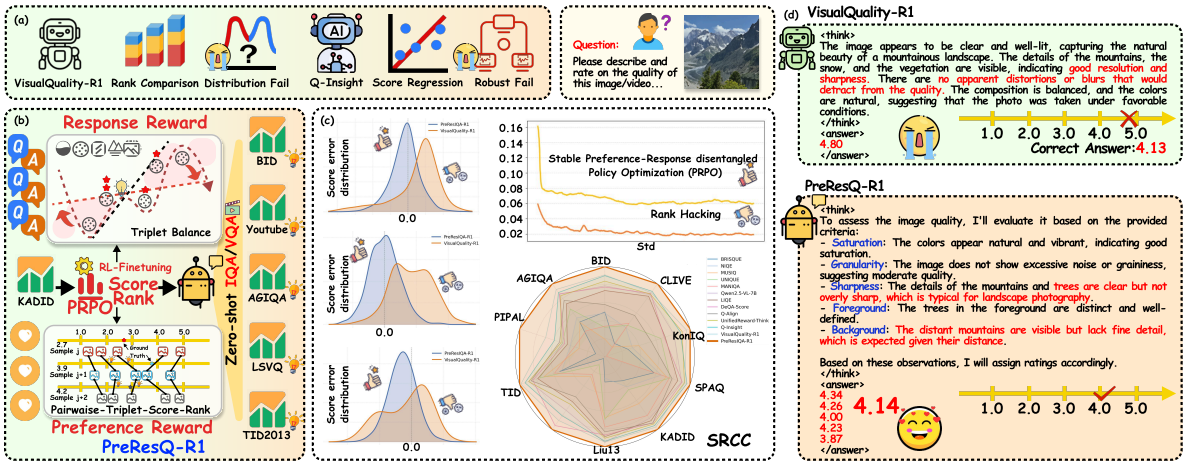


Figure 1: **Overview Performance**. (a) Existing score/ranking framework assign minimal difference, which results in distribution fall or robustness fail. (b) PreResQ-R1 focus on fine-grained response-ranking disentangled RL. (c) PreResQ-R1 enables state-of-the-art performance and stable image quality assessment. (d) typical comparisons with VisualQuality-R1 show that PreResQ-R1 achieves more accurate image quality description and scoring.

rather than exhaustive motion reconstruction (Feng et al., 2026a,b), highlighting the need for perceptually grounded evidence aggregation under practical input constraints (Zhang et al., 2025c; Ti et al., 2025; Wu et al., 2025b; Chen et al., 2026).

Motivated by these observations, we propose **PreResQ-R1**, a **Preference-Response Disentangled Reinforcement optimization** framework for visual QA. PreResQ-R1 decomposes perceptual learning into response-level stability and preference-level alignment, and integrates them within a unified Group Relative Policy Optimization (GRPO) scheme with long-chain reasoning. In particular, the response dimension is explicitly modeled to regularize stochastic generation and stabilize the learned score distribution, which in turn improves robustness and cross-distribution generalization, while the preference dimension enforces perceptual ordering with magnitude-aware calibration. For video QA task, we adopt a temporal-spatial evidence aggregation strategy that aligns with reasoning constraints without architectural modifications. Our contributions are threefold:

1. We propose a disentangled RL formulation that explicitly models response-level stability and preference-level alignment via complementary ranking-aware rewards.
2. We design a fine-grained reward scheme that jointly optimize ranking consistency and scoring alignment through complementary pairwise and triplet ranking signals.
3. Extensive experiments demonstrate that PreResQ-R1 achieves consistent improvement across diverse IQA and VQA benchmarks, producing interpretable, human-aligned reasoning traces.

2 Related Work

2.1 Regression-based models

Classical IQA and VQA methods predominantly formulate perceptual quality assessment as an absolute regression problem, mapping visual inputs to scalar scores based on signal fidelity assumptions or handcrafted natural scene statistics (Huynh-Thu and Ghanbari, 2008; Wang et al., 2004; Mittal et al., 2012, 2013). The emergence of deep learning enables end-to-end quality regression with CNN- and Transformer-based architectures, yielding substantial performance gains across benchmarks (Zhang et al., 2020; Ke et al., 2021; Wu et al., 2023a, 2022). Recent extensions to MLLMs incorporate pretrained multimodal representations to enhance semantic reasoning (Wu et al., 2023b; Li et al., 2025). However, these models entangle perceptual uncertainty with score prediction and lead to sensitivity to dataset-specific score distributions (Feng et al., 2026c, 2025).

2.2 Ranking-based models

Ranking-based IQA and VQA methods formulate quality assessment as a relative ranking task, which better aligns with human perceptual judgment. Early methods including RankIQA (Liu et al., 2017) optimize ranking objectives via correlation-based losses. Recent MLLM-based approaches, such as C2Score (Zhu et al., 2024), DeQA-Score (You et al., 2025), LMM-PVQA (Cao et al., 2025b), and VisualQuality-R1 (Wu et al., 2025a), further leverage comparative supervision to enhance both perceptual consistency and reasoning interpretability. However, ranking-based methods inherently sacrifice absolute score calibration,

making cross-dataset comparability challenging.

2.3 Reinforcement learning-based models

RL has recently emerged as a promising mechanism for aligning MLLMs with human perceptual preferences. Building on RLHF (Ouyang et al., 2022) and RLAIF (Yu et al., 2024), policy optimization frameworks such as GRPO (Shao et al., 2024) enable preference-driven refinement without explicit human annotations. This paradigm underlies models such as DeepSeek-R1-Zero (DeepSeek-AI, 2025), R1-VL (Zhang et al., 2025a), and Visual-RFT (Liu et al., 2025b). In the context of QA, RL-based methods aim to bridge perceptual fidelity and reasoning. Q-Insight (Li et al., 2025) and VQ-Insight (Zhang et al., 2025b) incorporate reward-based optimization for score prediction, while VisualQuality-R1 (Wu et al., 2025a) poses IQA as a reasoning-guided ranking problem. Related works on video and preference alignment (e.g. VideoDPO (Liu et al., 2024), VideoReward (Liu et al., 2025a), and UnifiedReward (Wang et al., 2025b)) demonstrate RL's potential, while also exposing challenges in optimization stability and computational efficiency.

3 Proposed Model

3.1 Overview

This work aims to incorporate long CoT reasoning into the reward MLLM's assessment process to enhance the robustness of reward signals. Evidence (Wu et al., 2025a, 2023b) shows that relying solely on score-based rewards often results in imbalanced score distributions, while rank-based rewards tend to be dataset-specific, leading to limitations in generalization across diverse tasks. The key challenge lies in devising an effective strategy to elicit and incentivize CoT capability. Therefore, we take the first step to proposed a preference-response disentangled multimodal CoT-based NR-QA model, PreResQ-R1 (Fig. 2), to activates both score and rank convergence through fine-grained reasoning.

3.2 Fine-grained CoT for IQA and VQA

Given a text prompt $\mathcal{P}$ and an image or video $\mathcal{I}$ in an image/video, our goal is to fine-tune a pretrained MLLM with parameters θ and policy $\pi_\theta(\cdot|\mathcal{P},\mathcal{I})$ to generate quality reasoning trajectories τ that provide a comprehensive CoT c along with an overall score $s \in [1,5]$:

$$\tau = (c, s) \sim \pi_\theta(\cdot|\mathcal{P},\mathcal{I}) \tag{1}$$

We present the complete structured text prompt in Appendix H. For each image-score pair $x_i = (\mathcal{P}_i, \mathcal{I}_i)$ in IQA, the MLLM generates detailed Chain-of-Thought (CoT) reasoning for five quality aspects: *saturation*, *granularity*, and *sharpness*, which capture pixel-level properties, and *foreground* and *background*, which capture semantic-level information, followed by the corresponding numeric scores. For VQA, inspired by DeepSeek-OCR (Wei et al., 2025), we adopt a temporal-spatial representation compatible with IQA task. We uniformly sample M frames to form a temporal-flow image summarizing long-range temporal evolution, and N frames to construct local spatial-flow images capturing frame-level details. Although this formulation does not explicitly model fine-grained temporal continuity or motion-specific artifacts, it effectively exploits perceptual priors learned from IQA fine-tuning and enables scalable VQA under multilevel linear MLLM input constraints. Consequently, VQA is performed via structured aggregation of temporally distributed visual evidence rather than explicit motion reconstruction. For every training batch of image/video-score pairs $\{x_i, x_{i+1}, ..., x_{i+B}\}$, where B is the minibatch size, we apply GRPO to generate K quality predictions for x_i:

$$\tau(x_i) = [\tau_1(x_i), \tau_2(x_i), ..., \tau_K(x_i)] \tag{2}$$

3.3 Preference-Response disentangled Policy Optimization (PRPO)

PRPO explicitly decomposes the policy optimization into two complementary components that jointly capture the essence of perceptual reasoning. The **Response Dimension** focuses on intra-sample consistency, evaluating how coherently the model reasons across multiple generations of the same sample. Meanwhile, the **Preference Dimension** models inter-sample relations, encouraging the predicted quality score to align with human MOS. In addition, a format reward R_{for} is utilized to ensure specific answer format.

3.3.1 Response-based ranking reward

To accelerate early-stage convergence and avoid degenerate score collapse caused by highly implausible predictions, we introduce a response-based ranking reward that imposes a weak intra-sample consistency bias across stochastic generations. For each input sample j, given a triplet of stochastic predictions $\{s^l_{j,d}, s^m_{j,d}, s^n_{j,d}\}$ on perceptual dimension d, we use their median as a local reference and

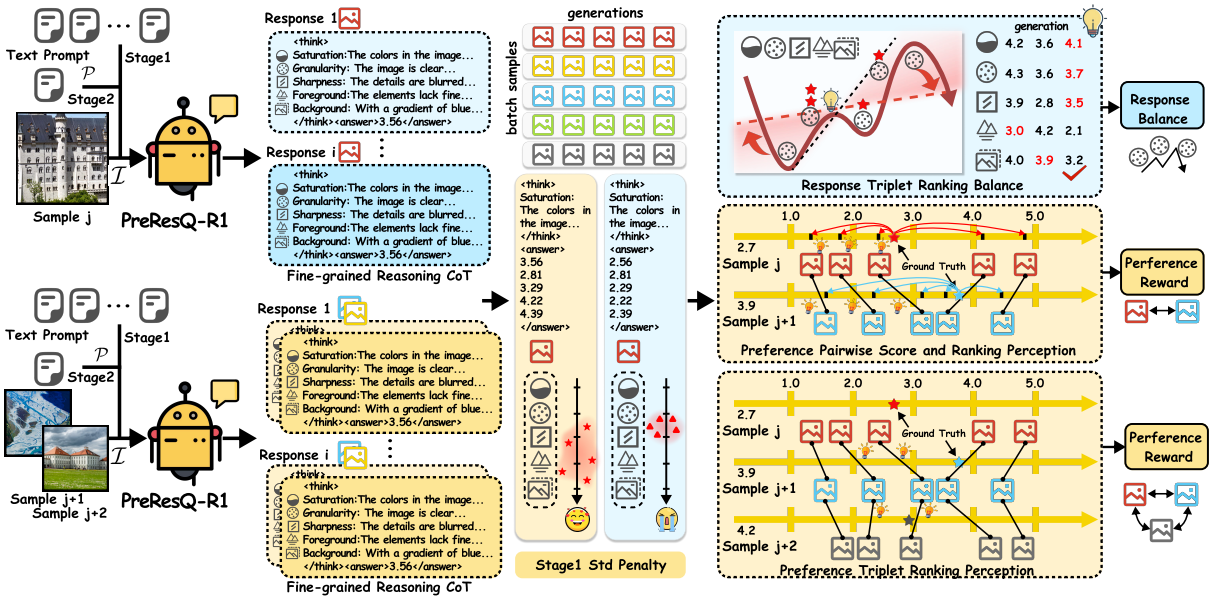


Figure 2: Overall training framework of PreResQ-R1 via reinforcement learning to rank and score. Given an sample batch ($\mathcal{I}_j$, $\mathcal{I}_{j+1}$,..., $\mathcal{I}_{j+B}$) with a shared text prompt $\mathcal{P}$, PreResQ-R1 generates K responses. To quickly activate CoT differences and then access generation stability, we introduce the response penalty and fine-grained triplet-response balance reward. To jointly enhance the robustness of ranking and score ability, we introduce the preference pairwise-and-triplet score-and-ranking reward for GRPO.

define the response-based ranking reward as:

$$R_{RR}^{j,i} = \frac{1}{D}\sum_{d=1}^{D}\exp\Big(-\Big|s_{j,d}^{i} - \text{median}\Big(\{s_{j,d}^{l}, s_{j,d}^{m}, s_{j,d}^{n}\}\Big)\Big|\Big), \tag{3}$$

This formulation suppresses extreme outliers while preserving meaningful score variability. To ensure numerical stability, triplets with a total absolute difference below 0.5 receive a reward of 1. RR serves as an early-phase optimization scaffold, guiding exploration away from implausible score regions, and naturally becomes secondary to preference-driven rewards as policy entropy decreases.

3.3.2 Preference-based ranking-score reward

Pairwise Comparative Reward.
Building upon the response-based ranking reward, which constrains stochastic generations of the same input into a stabilized response distribution, we further introduce a preference-based reward to enable effective inter-sample quality comparison. Under the RR-induced stabilization, multiple responses to the same input exhibit reduced variance and comparable relative ordering, providing a prerequisite for preference learning. In quality assessment, purely score-based rewards are prone to distribution imbalance and score collapse, while rank-based rewards, although more robust to scale variations, often suffer from dataset-dependent ordinal bias and limited sensitivity to perceptual gaps. To balance these complementary strengths and weaknesses, we adopt a rank-score coupled preference formulation that jointly enforces ordinal correctness and magnitude awareness. Specifically, for each sample j, we first perform an intra-sample ranking over its K stochastic generations based on their predicted quality scores, yielding ordered statistics $\hat{s}_j^{(1)} \le \hat{s}_j^{(2)} \le \cdots \le \hat{s}_j^{(K)}$. This ranking serves as a normalization step that preserves relative quality structure learned under RR. Cross-sample preference learning is then conducted by comparing predictions at the same rank index i across different samples. Rather than assuming absolute score comparability, this design leverages the RR-induced convergence property: predictions at the same order statistic index represent similar relative confidence levels across samples, enabling stable preference supervision.

Formally, let $\mathbf{G}_j$ denote the ground-truth MOS for sample j. For any sample pair (l, m) and rank index i, we define an ordinal consistency indicator $\mathbb{I}$:

$$C_{l,m}^{(i)} = \mathbb{I}\Big[\text{sign}\big(\hat{s}_l^{(i)} - \hat{s}_m^{(i)}\big) = \text{sign}\big(\mathbf{G}_l - \mathbf{G}_m\big)\Big], \tag{4}$$

which measures whether the i-th ranked predictions preserve the ground-truth ordering. To complement ordinal supervision with perceptual sensitivity, we introduce a magnitude-aware alignment term:

$$M_{l,m}^{(i)} = \frac{|\mathbf{G}_l - \mathbf{G}_m|}{|\hat{s}_l^{(i)} - \hat{s}_m^{(i)}| + |\hat{s}_l^{(i)} - \mathbf{G}_l| + |\hat{s}_m^{(i)} - \mathbf{G}_m|}, \tag{5}$$

The final preference-based ranking-score reward for generation i of sample l in minibatch size B is defined as:

$$R_{pair}^{l,i} = \sum_{\substack{m=1 \\ m \neq l}}^{B} \left[\sqrt{C_{l,m}^{(i)} e^{M_{l,m}^{(i)}}} + \sqrt{(1 - C_{l,m}^{(i)}) e^{-(1+M_{l,m}^{(i)})}} \right]. \tag{6}$$

Triplet comparative reward: Pairwise ranking supervision enforces local ordinal correctness but does not guarantee global ranking consistency. In particular, pairwise-consistent predictions may still violate transitivity, leading to unstable or contradictory preference structures. To complement pairwise constraints, we introduce a triplet-based comparative reward that explicitly enforces transitive consistency among three samples. By jointly evaluating all three pairwise relations within a triplet, the triplet reward regularizes the global ranking geometry and suppresses cyclic inconsistencies that cannot be resolved by pairwise supervision alone. The pairwise and triplet rewards thus serve complementary roles: the former provides fine-grained local preference signals, while the latter acts as a structural constraint that stabilizes global ordering. We adopt a simple ordinal triplet formulation that avoids additional scale assumptions and preserves training stability. For each anchor sample a, we construct the set of all triplets $\mathcal{T}_l = \{(l, m, n) \,|\, m, n \neq l\}$. For a single triplet $(l, m, n) \in \mathcal{T}_i$, the triplet reward is defined as:

$$r_{tri}^{(l,m,n),i} = \begin{cases} 1.0, & \text{if } C_{l,m}^{(i)} + C_{l,n}^{(i)} + C_{m,n}^{(i)} = 3, \\ 0.3, & \text{otherwise,} \end{cases} \tag{7}$$

where $C_{l,m}^{i} \in \{0, 1\}$ denotes the pairwise consistency between predictions s_l^k, s_m^k and ground-truth scores $\mathbf{G}_l, \mathbf{G}_m$. The overall triplet reward for generation i and sample j is then computed as the mean over all triplets:

$$R_{tri}^{j,i} = \frac{1}{|\mathcal{T}_j|} \sum_{(l,m,n) \in \mathcal{T}_j} r_{tri}^{(l,m,n),i}. \tag{8}$$

3.4 Total objective

We unify response-level fidelity and cross-sample preference supervision into a single scalar reward. For the i-th generation of sample j, the total reward is defined as

$$R_{\text{total}}^{j,i} = R_{\text{for}}^{j,i} + \alpha\, R_{\text{RR}}^{j,i} + (1 - \alpha)\big(\beta\, R_{\text{pair}}^{j,i} + (1 - \beta)\, R_{\text{tri}}^{j,i}\big), \tag{9}$$

where $\alpha, \beta \in [0, 1]$. This hierarchical design constrains individual responses to be perceptually faithful and distributionally stable while enforcing both local and global preference consistency across samples. The resulting objective aligns intra-response quality with inter-sample ordering, enabling stable reinforcement fine-tuning under stochastic generation. The model is optimized by a regularized policy gradient objective. Detailed reinforcement learning formulations are provided in Appendix I.

3.5 Exploration-to-stability fine-tuning strategy

To balance exploration and stable fine-tuning, we adopt a two-stage fine-tuning strategy. In the exploration stage (τ), each sample is paired with $\mathbf{X}$ randomly sampled prompts, encouraging the model to generate diverse trajectories. For each generation, we extract the five-dimensional scores $\mathbf{s}_j^{(i)}$ and compute a dimension-wise standard deviation penalty:

$$R_{\text{std}}(\mathbf{s}_j^{(i)}) = \begin{cases} \lambda_{\text{std}}(\delta_{\min} - \sigma(\mathbf{s}_j^{(i)}), & \sigma(\mathbf{s}_j^{(i)}) < \delta_{\min}, \\ 0, & \text{otherwise,} \end{cases} \tag{10}$$

where $\mathbf{s}_j^{(i)} = [s_j^1, \ldots, s_j^5]$ denotes the five-dimensional quality scores for the i-th generation of sample j, and $\sigma(\mathbf{s}_j^{(i)})$ is the standard deviation across these five scores.

4 Experiment

4.1 Experimental setups

For **IQA task**, PreResQ-R1 is trained exclusively on the synthetic KADID-10K and KonIQ dataset (6:2:2 split in content independence) and tested in a zero-shot setting across ten open-source general (real-world and AI-generated) image aesthetic assessment datasets: BID (Ciancio et al., 2011), CLIVE (Ghadiyaram and Bovik, 2016), KonIQ-10k (Hosu et al., 2020), SPAQ (Fang et al., 2020), KADID (Lin et al., 2019), Liu 13 (Liu et al., 2013), SRIQA-Bench (Chen et al., 2025a), AGIQA-3K (Li et al., 2023), TID 2013 (Ponomarenko et al., 2015), and PIPAL (Gu et al., 2020).

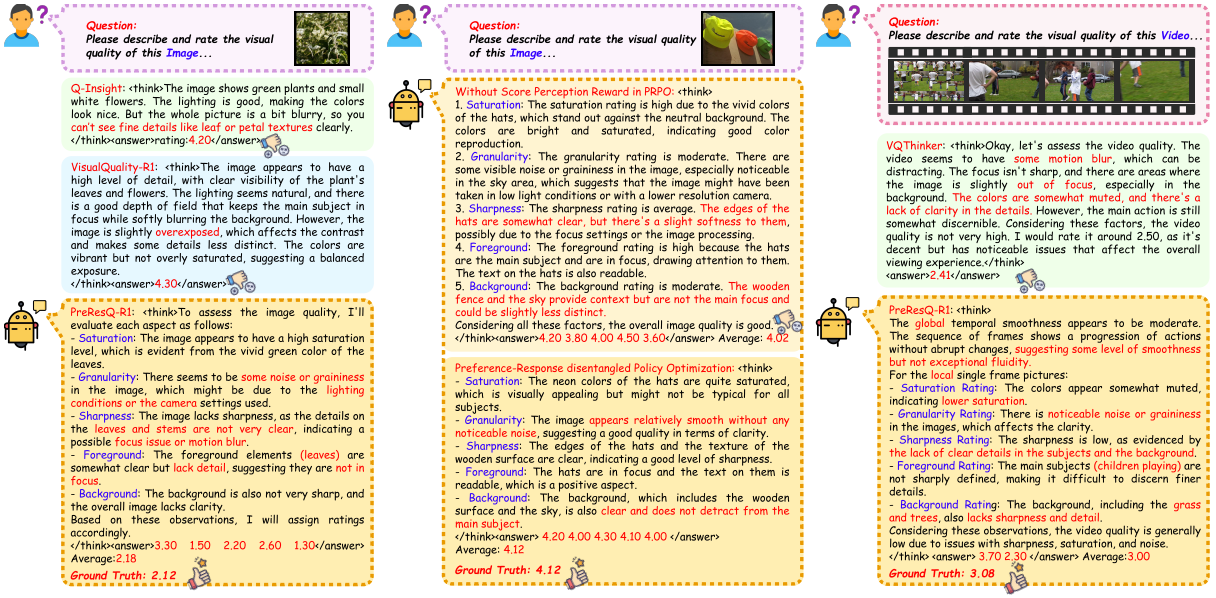
For **VQA task**, PreResQ-R1 is trained exclusively on the LSVQ-28K (Ying et al., 2021) dataset and test in a zero-shot setting across five open-source in-domain UGC videos with in-the-wild distortions: $\text{LSVQ}_{\text{test}}$ (Ying et al., 2021), KoNViD-1k (Hosu et al., 2020), LIVE-VQC (Sinno and Bovik, 2019), YouTube-UGC(Wang et al., 2019).

4.2 Implementation details

We conduct IQA PreResQ-R1 in two stages. In the first stage, prompts are randomly selected

Table 1: Performance comparison of PreResQ-R1 across KADID-10K IQA datasets. For each dataset, SRCC (S) and PLCC (P) are reported. **Red**, **green**, and **blue** denote handcrafted, deep-learning-based, and MLLM-based methods. **Bold** and underline indicate the best and second best results. All reported results are obtained by averaging over ten independent runs with different random seeds.

Method	BID		CLIVE		KonIQ		SPAQ		KADID		Liu13		SRIQA		TID13		PIPAL		AGIQA		Avg	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
BRISQUE	0.522	0.528	0.314	0.362	0.385	0.400	0.614	0.624	0.429	0.451	0.389	0.380	0.210	0.163	0.548	0.546	0.242	0.259	0.497	0.541	0.415	0.425
NIQE	0.515	0.527	0.450	0.494	0.421	0.439	0.676	0.683	0.487	0.415	0.360	0.376	0.194	0.135	0.532	0.516	0.357	0.314	0.533	0.560	0.453	0.446
MUSIQ	0.327	0.280	0.284	0.325	0.473	0.435	0.720	0.666	0.647	0.622	0.656	0.563	0.309	0.283	0.670	0.695	0.317	0.347	0.494	0.434	0.490	0.456
UNIQUE	0.412	0.385	0.470	0.472	0.649	0.590	0.751	0.708	0.513	0.548	0.669	0.654	0.244	0.207	0.703	0.729	0.393	0.361	0.608	0.581	0.541	0.524
MANIQA	0.420	0.512	0.487	0.571	0.213	0.257	0.745	0.753	0.760	0.780	0.726	0.728	0.297	0.289	0.589	0.617	0.338	0.364	0.422	0.448	0.500	0.532
Qwen2.5-VL	0.711	0.725	0.733	0.760	0.754	0.810	0.848	0.854	0.787	0.806	0.840	0.852	0.495	0.504	0.787	0.837	0.390	0.420	0.735	0.772	0.708	0.734
LIQE	0.677	0.680	0.719	0.726	0.684	0.652	0.815	0.814	0.809	0.817	0.797	0.712	0.447	0.430	0.718	0.748	0.371	0.347	0.653	0.653	0.669	0.658
DeQA-Score	0.702	0.743	0.743	0.795	0.677	0.703	0.852	0.858	0.831	0.873	0.785	0.838	0.579	0.589	0.756	0.793	0.383	0.381	0.738	0.743	0.705	0.732
Q-Align	0.576	0.651	0.554	0.643	0.573	0.612	0.767	0.779	0.832	0.862	0.761	0.802	0.421	0.459	0.769	0.794	0.406	0.381	0.682	0.694	0.634	0.668
UnifiedReward-T	0.745	0.767	0.738	0.795	0.820	0.804	0.871	0.846	0.841	0.877	0.809	0.835	0.462	0.518	0.788	0.873	0.399	0.421	0.722	0.745	0.720	0.748
Q-Insight	0.784	0.796	0.761	0.795	0.806	0.829	0.892	0.872	0.856	0.881	0.831	0.857	0.525	0.542	0.816	0.851	0.429	**0.485**	0.749	0.794	0.745	0.770
VisualQuality-R1	0.790	0.806	0.750	0.794	0.855	0.840	0.875	0.878	0.871	0.821	0.838	**0.872**	0.561	0.543	0.848	0.871	0.469	0.458	0.805	**0.843**	0.768	0.773
PreResQ-R1 (Ours)	**0.849**	**0.839**	**0.825**	**0.825**	**0.884**	**0.887**	**0.919**	**0.900**	**0.894**	**0.892**	**0.873**	0.864	**0.601**	**0.596**	**0.858**	**0.877**	**0.588**	0.431	**0.823**	0.809	**0.811**	**0.790**



(a) Image Quality Assessment Case

(a) Video Quality Assessment Case

Figure 3: (a) Qualitative cases of IQA in comparison and ablation studies. (b) Qualitative cases of VQA in comparison. Fine-grained perceptual attributes are highlighted in purple, and quality-relevant visual evidence in red.

from a predefined set of $\mathbf{X} = 5$, with standard deviation (std) penalties set as $\delta_{\min} = 0.5$ and $\lambda_{\text{std}} = 0.5$, and generate $K = 12$ responses each sample. In the second stage, we generate $K = 6$ responses each sample. We set α=0.5, β=0.75. We fine-tune Qwen2.5-VL-7B (Bai et al., 2025) with GRPO (Shao et al., 2024) without freezing any layers. For VQA task, we set $K = 5$, $M = 12$, and $N = 3$. PreResVQA-R1 is fine-tuned based on the PreResIQA-R1 only in single-prompt stage. Training is performed using AdamW (Loshchilov and Hutter, 2019) with an initial learning rate of 3×10^{-7}, a linear decay schedule. All experiments are conducted on 8 NVIDIA A800 GPUs, with a per-GPU batch size of 48 for IQA and 25 for VQA. The models are trained for 8 epochs on IQA and 1 epoch on VQA, requiring approximately 20 hours and 35 hours of total training time, respectively.

4.3 Comparison studies

For **IQA** task, we perform quantitative and qualitative comparisons with handcrafted, deep learning-based, and MLLM-based methods, as summarized in Table 1. When trained on KADID-10K, PreResQ-R1 achieves an average SRCC of **0.811** and PLCC of **0.795** across 10 IQA benchmarks, surpassing prior approaches by **5.6%** and **2.2%**, respectively. Notably, PreResQ-R1 attains the highest SRCC on all evaluated datasets, indicating consistently improved perceptual ordering enabled by the proposed pairwise-triplet score-ranking reward. Furthermore, we also provide quantitative results trained on KonIQ dataset (Appendix E).

Qualitative comparisons in Fig. 3(a) show that PreResQ-R1 produces reasoning traces that are tightly grounded in visual evidence, with consistent attribution of quality-relevant factors. In

Table 2: Performance comparison of PreResQ-R1 trained on LSVQ-28K dataset and evaluated across different VQA datasets under SRCC and PLCC metrics. **Red**, **green**, and **blue** denote Handcrafted, Discriminative deep-learning, and MLLM-based methods. **Bold** and underlined indicate the best and second best results. ◁ denotes trained on fused-287K dataset, ◇ denotes trained on LSVQ-28K dataset, ⋆ denotes fine-tuned based on PreResIQA.

Method	$LSVQ_{test}$		KoNViD-1k		LIVE-VQC		YouTube-UGC		Avg	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
STEM◇	0.206	0.243	0.619	0.627	0.594	0.629	0.284	0.318	0.426	0.454
NIQE◇	0.442	0.332	0.541	0.553	0.596	0.628	0.278	0.290	0.464	0.451
FAST-VQA◇	0.880	0.880	0.859	0.854	0.826	0.845	0.730	0.747	0.824	0.832
MinimalisticVQA◇	0.885	0.882	0.862	0.859	0.775	0.821	0.826	0.821	0.837	0.846
VQ-Insight◁ (no code)	0.875	0.876	0.875	0.884	0.790	0.835	–	–	–	–
Q-Align◇	0.886	0.884	0.876	0.878	0.783	0.819	0.834	0.846	0.845	0.857
VQA^2-Scorer◇	0.878	0.872	0.881	0.880	0.785	0.830	**0.811**	0.823	0.839	0.851
Q-Insight◇	0.644	0.639	0.751	0.753	0.624	0.708	0.560	0.591	0.645	0.673
VisualQuality-R1◇	0.795	0.796	0.784	0.792	0.732	0.781	0.717	0.730	0.757	0.576
PreResQ-R1 (Ours)◇	0.865	0.853	0.843	0.851	0.774	0.762	0.795	0.806	0.819	0.818
VQAThinker◇	0.883	0.880	0.881	0.884	0.808	0.847	0.860	0.863	0.858	0.869
PreResQ-R1 (Ours)⋆	**0.901**	**0.893**	**0.879**	**0.889**	**0.838**	**0.895**	**0.870**	**0.887**	**0.872**	**0.891**

contrast, existing methods such as Q-Insight and VisualQuality-R1 may produce fluent explanations that are not consistently supported by visual evidence. By explicitly constraining response consistency and preference alignment, PreResQ-R1 encourages evidence-aligned CoT reasoning, resulting in improved interpretability and reliability.

For **VQA** task, we compare PreResQ-R1 with comparable VQA methods, as reported in Table 2. PreResQ-R1 achieves an average SRCC of **0.872** and PLCC of **0.891** across 4 datasets. As shown in Fig. 3(b), compared with VQ-Thinker, PreResQ-R1 yields more accurate predictions accompanied by structured reasoning that jointly captures global temporal coherence, local spatial distortions, and fine-grained attributes. Notably, PreResVQA-R1 is fine-tuned from PreResIQA-R1, demonstrating that perceptual priors learned from IQA can be effectively transferred to VQA. This design bridges IQA and VQA without introducing additional video-specific encoding parameters, enabling parameter-efficient adaptation while preserving strong reasoning capability under practical input constraints (Appendix H).

4.4 Ablation studies

We further analyze the impact of **individual reward components** in Table 3. PSR establishes a stable baseline, confirming that pairwise preference learning is effective yet insufficient for capturing global perceptual consistency. Incorporating RTR brings a consistent performance gain of around 3%, highlighting the importance of constraining intra-sample response variability to stabilize reasoning and score generation. Introducing PTR yields further improvements of further 3% by enforcing globally higher-order perceptual ordering beyond pairwise relations. The complementary effects of Response-dim and Preference-dim (Tab. 6) are also detailed in Appendix B. Moreover, we provide additional **absolute-score-relative-rank ablation qualitative studies** on the IQA task in Fig.3 (a) and temporal-spatial analyses on the VQA task (Appendix H). Without the score perception reward, the reasoning CoT becomes less coherent and the score degrades significantly, demonstrating the rationality of the disentangled RL design.

Furthermore, we conduct **computational cost ablations** in Table 4. For **IQA** task, incorporating additional reward components doesn't measurably affect training cost, with inference time at 0.6 s/sample. For **VQA** task, the TSF strategy introduces additional cost mainly from image encoding; however, it remains substantially lighter than full video encoding, with inference time at 2.95 s/sample. These results indicate that PreResQ-R1 doesn't introduce significant computational burden, validating its practical efficiency.

Table 3: Performance ablation of PreResQ-R1 across different reward functions. RTR, PSR, and PTR represent the Response Triplet Ranking reward, Preference Pairwise Score-Ranking reward, and Preference Triplet Ranking reward, respectively. No performance collapse is observed under hyper-parameter variations.

RTR	PSR	PTR	BID		KonIQ		SPAQ		TID2013		AGIQA	
			PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
✗	✓	✗	0.782	0.767	0.839	0.859	0.844	0.834	0.775	0.796	0.762	0.759
✗	✗	✓	0.765	0.790	0.868	0.875	0.858	0.829	0.794	0.781	0.770	0.764
✓	✓	✗	0.795	0.813	0.872	0.867	0.882	0.890	0.823	0.847	0.791	0.808
✓	✗	✓	0.816	0.809	**0.895**	**0.904**	0.896	0.889	0.842	0.858	0.785	0.779
✓	✓	✓	**0.849**	**0.839**	0.884	0.887	**0.919**	**0.900**	**0.858**	**0.877**	**0.823**	**0.809**

Table 4: Computational cost ablation of PreResQ-R1 across different reward functions (RTR, PSR, and PTR functions). TSF represents the Temporal-Spatial Flow strategy for VQA task.

RTR	PSR	PTR	TSF	Training Memory	Training Time	Test Memory	Test Time
✗	✗	✓	✗	68.53±2.94 GB	2.09±0.12 h/10K samples	33.32±0.72 GB	0.52 s/sample
✗	✓	✗	✗	68.93±1.58 GB	2.12±0.09 h/10K samples	33.51±0.75 GB	0.41 s/sample
✓	✓	✗	✗	70.03±1.62 GB	2.17±0.11 h/10K samples	33.35±0.85 GB	0.69 s/sample
✓	✗	✓	✗	70.26±1.69 GB	2.26±0.15 h/10K samples	33.26±0.79 GB	0.60 s/sample
✓	✓	✓	✗	71.28±2.03 GB	2.38±0.23 h/10K samples	33.61±0.83 GB	0.72 s/sample
✓	✓	✓	✓	73.37±2.91 GB	10.21±0.38 h/10K samples	32.63±0.70 GB	2.95 s/sample
VQA-Thinker				–	–	42.62±0.61 GB	3.83 s/sample

Table 5: Quantitative evaluation on the FineVD dataset (Duan et al., 2025) across five distortion categories (Color, Noise, Artifact, Blur, Temporal) in terms of SRCC and PLCC. Our method PreResQ consistently outperforms existing baselines including Q-Align, VisualQuality, and VQAThinker on all distortion types.

Model	Color		Noise		Artifact		Blur		Temporal	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
Q-Align	0.529	0.554	0.506	0.553	0.554	0.569	0.591	0.582	0.505	0.512
VisualQuality	0.595	0.609	0.584	0.567	0.632	0.658	0.637	0.640	0.516	0.529
VQAThinker	0.620	0.657	0.608	0.605	0.628	0.634	0.629	0.635	0.539	0.561
PreResQ	0.655	0.672	0.635	0.652	0.664	0.660	0.665	0.652	0.595	0.573

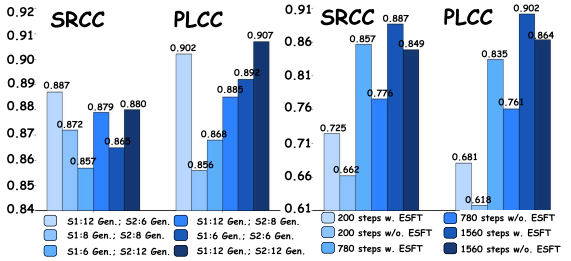


Figure 4: Left: results on 10 synthetic evaluation datasets (each dataset 100 images) with different generation number $Gen.$ across two stages $S1/S2$; right: results on evaluation datasets with/without Exploration-to-Stability Fine-Tuning (ESFT).

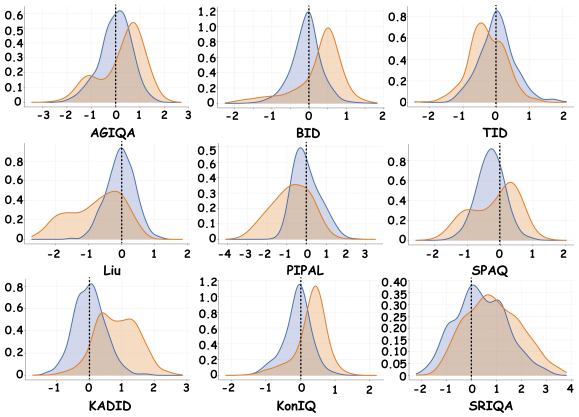


Figure 5: Comparison between PreResQ-R1 and VisualQuality-R1 (Wu et al., 2025a) on distribution of difference between answer and ground truth. The horizontal axis represents the error, and the vertical axis represents the relative proportion. The closer the distribution is to 0, the better the model performance is. Blue and Orange denotes PreResQ-R1 and VisualQuality-R1.

Finally, we analyze the **ablation of response generation** and the **ESFT training dynamics** on synthetic IQA dataset. Fig. 4 left shows that increasing response diversity in Stage 1 is crucial for effective exploration, while a balanced generation-comparison scheme in Stage 2 stabilizes preference learning, jointly leading to optimal performance. As shown in Fig. 4 right, ESS significantly accelerates convergence and enhances multi-granularity prediction accuracy by smoothly transitioning from exploration to stable optimization.

4.5 Comprehensive Studies

Fig. 6 visualizes the **training dynamics** of PreResQ-R1, including reward standard deviation (std), fine-grained CoT answer std, response generation std. As shown in Fig. 6(a), the reward std demonstrates that the disentangled RL formulation moves beyond a single static target, progressively fitting discriminative pairwise and triplet preferences. Fig. 6(b) shows a low-to-high evo-

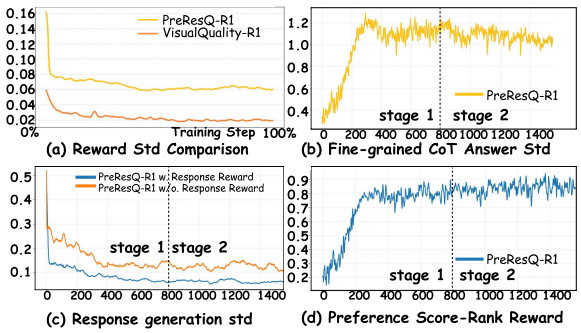


Figure 6: (a). Reward standard deviation comparison between VisualQuality-R1 and PreResQ-R1. (b). Fine-grained CoT intra answer score standard deviation. (c). generation standard deviation comparison. (d). Total reward tendency visualization.

lution of CoT standard deviation, followed by a slight reduction in Stage 2, suggesting that the std penalty functions as an early exploration scaffold and diminishes as preference-driven rewards take over with reduced policy entropy. Introducing the RTR reward further stabilizes response generation (Fig. 6(c)), constraining intra-sample variability and enabling consistent and faster convergence.

Finally, Fig. 6(d) illustrates the progressive stabilization and improvement of the preference score-rank reward, reflecting the model's enhanced capability to align inter-sample preferences while preserving magnitude-sensitive calibration. More hyperparamter scanning results are also available in Appendix A. Finally, we provide detailed evaluations of GPT-5 mini reasoning, the contributions of the PRPO and TSF modules, and a comprehensive analysis of perceptual-level performance across fine-grained IQA and VQA attributes (Appendices C, D, and F). These results collectively offer empirical support for the design of complementary response- and preference-level objectives, illustrating how each component contributes to improved training stability, enhanced reasoning fidelity, and consistent global perceptual alignment.

5 Limitaion and future work

Despite its effectiveness, PreResQ-R1 has several limitations. First, the disentangled formulation assumes that response stability and preference alignment can be cleanly separated, which may not fully capture their inherent coupling in human perceptual judgments. Second, while response-level regularization improves robustness, it does not fully resolve uncertainty under highly subjective or out-of-distribution conditions, where quality perception lacks a well-defined consensus. Third, the

proposed video extension relies on sparse temporal evidence and may be insufficient for scenarios requiring fine-grained motion reasoning. Future work will explore more principled formulations of perceptual uncertainty, including unified models that jointly capture stability and preference under stochastic generation. Extending the framework to long-form and dynamic visual content with richer temporal modeling is another direction. Improving robustness under distribution shifts and incorporating adaptive or human-in-the-loop feedback may further enhance reliability.

6 Conclusion

We introduced PreResQ-R1, a preference-response disentangled reinforcement optimization framework that unifies score regression and ranking consistency. With only 6K image samples and 28K video samples, it achieves comparable performance across 10 IQA datasets and 5 VQA datasets while providing reasonable CoT. Future work will aim to integrate image quality assessment with text-to-image generation within a unified framework, enabling complementary optimization.

Acknowledgments

This work is partly supported by National Natural Science Foundation of China under Grant No. 72471140, partly by Shanghai Jiao Tong University Key Projects of Smart Liberal Arts under Grant No. ZHWK2505, partly by Fundamental Research Funds for the Central Universities under Grant No. YG2025QNA13, partly by Shanghai Magnolia Talent Program-Pujiang Project under Grant No. 2025PJC167, partly by the Lushan Research Funding.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Xin Cai, Zhiyuan You, Hailong Zhang, Wentao Liu, Jinwei Gu, and Tianfan Xue. 2024. Phocolens: Photorealistic and consistent reconstruction in lensless imaging. *arXiv preprint arXiv:2409.17996*.

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

Linhan Cao, Wei Sun, Xiangyang Zhu, Kaiwei Zhang, Jun Jia, Yicong Peng, Dandan Zhu, Guangtao Zhai, and Xiongkuo Min. 2025b. Towards generalized video quality assessment: A weak-to-strong learning paradigm. *arXiv preprint arXiv:2505.03631*.

Du Chen, Tianhe Wu, Kede Ma, and Lei Zhang. 2025a. Toward generalized image quality assessment: Relaxing the perfect reference quality assumption. *arXiv preprint arXiv:2503.11221*.

Jianan Chen, Zhifang Zhang, Shuo He, Linan Yue, Lei Feng, and Minling Zhang. 2026. Towards safer large reasoning models by promoting safety decision-making before chain-of-thought generation. *arXiv preprint arXiv:2603.17368*.

Zijian Chen, Wei Sun, Haoning Wu, Zicheng Zhang, Jun Jia, Ru Huang, Xiongkuo Min, Guangtao Zhai, and Wenjun Zhang. 2025b. Study of subjective and objective naturalness assessment of ai-generated images. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(4):3573–3588.

Zixuan Chen, Zewei He, Ziqian Lu, Xuecheng Sun, and Zhe-Ming Lu. 2024. Prompt-based test-time real image dehazing: a novel pipeline. *arXiv preprint arXiv:2309.17389*.

Alexandre Ciancio, André Luiz N Targino Targino da Costa, Eduardo A. B. da Silva, Amir Said, Ramin Samadani, and Pere Obrador. 2011. No-reference blur assessment of digital pictures based on multifeature classifiers. *IEEE Transactions on Image Processing*, 20(1):64–75.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Huiyu Duan, Qiang Hu, Jiarui Wang, Liu Yang, Zitong Xu, Lu Liu, Xiongkuo Min, Chunlei Cai, Tianxiao Ye, Xiaoyun Zhang, and Guangtao Zhai. 2025. Finevq: Fine-grained user generated content video quality assessment.

Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. 2020. Perceptual quality assessment of smartphone photography. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3674–3683.

Zehui Feng, Tong Wu, Kefan Wu, Zhaopeng Zhu, Huayuan Xu, and Ting Han. 2026a. Triemo: Triple semantic alignment based on modality pair contrastive learning graph network for multimodal emotion recognition. *Information Fusion*, 133:104289.

Zehui Feng, Tong Wu, Dian Zhu, Susana Liu, Hongtao Zheng, Huiwu Li, and Ting Han. 2026b. A novel

multi-graph-structure guided human lower-limb physiological load estimation method based on multi-inertial-sensors fusion during walking. *Information Fusion*, 127:103914.

Zehui Feng, Tong Wu, Zhaopeng Zhu, and Ting Han. 2026c. Hypergraph-based global–local virtual unified emotion fusion for multimodal emotion recognition in conversations. *Pattern Recognition*, 180:114584.

Zehui Feng, Yangge Yang, Chenqi Zhang, Junxuan Li, and Ting Han. 2025. A novel hierarchical attention-guided refinement method with eeg assistance for enhancing target speech in a multi-speaker competing environment. *Advanced Engineering Informatics*, 65:103363.

Deepti Ghadiyaram and Alan C. Bovik. 2016. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387.

Jinjin Gu, Haoming Cai, Haoyu Chen, Xiaoxing Ye, Jimmy Ren, and Chao Dong. 2020. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. *arXiv preprint arXiv:2007.12142*.

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

Q. Huynh-Thu and M. Ghanbari. 2008. Scope of validity of psnr in image/video quality assessment. *Electronics Letters*, 44:800–801.

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

Parimala Kancharla and Sumohana S. Channappayya. 2022. Completely blind quality assessment of user generated video content. *IEEE Transactions on Image Processing*, 31:263–274.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. 2021. Musiq: Multi-scale image quality transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5128–5137.

Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. 2023. Agiqa-3k: An open database for ai-generated image quality assessment. *arXiv preprint arXiv:2306.04717*.

Hao Li, Yuhao Wang, Wenning Hao, Pingping Zhang, Dong Wang, and Huchuan Lu. 2026a. Ragtrack: Language-aware rgbt tracking with retrieval-augmented generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28179–28189.

Hao Li, Yuhao Wang, Xiantao Hu, Wenning Hao, Pingping Zhang, Dong Wang, and Huchuan Lu. 2026b. Cadtrack: Learning contextual aggregation with deformable alignment for robust rgbt tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 6109–6117.

Quanjiang Li, Zhiming Liu, Wei Luo, Tingjin Luo, and Chenping Hou. 2026c. Correcting visual blur induced by attention distraction to reduce hallucinations: Algorithm and theory. *arXiv preprint arXiv:2605.24602*.

Quanjiang Li, Tingjin Luo, Mingdie Jiang, Jiahui Liao, and Zhangqi Jiang. 2024. Deep incomplete multi-view network semi-supervised multi-label learning with unbiased loss. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9048–9056.

Quanjiang Li, Tianxiang Xu, Tingjin Luo, Yan Zhong, Yang Li, Yiyun Zhou, and Chenping Hou. 2026d. Theory-driven label-specific representation for incomplete multi-view multi-label learning. *Advances in Neural Information Processing Systems*, 38:121996–122037.

Weiqi Li, Xuanyu Zhang, Shijie Zhao, Yabin Zhang, Junlin Li, Li Zhang, and Jian Zhang. 2025. Q-insight: Understanding image quality via visual reinforcement learning. *arXiv preprint arXiv:2503.22679*.

Hanhe Lin, Vlad Hosu, and Dietmar Saupe. 2019. Kadid-10k: A large-scale artificially distorted iqa database. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, Xintao Wang, Xiaohong Liu, Fei Yang, Pengfei Wan, Di Zhang, Kun Gai, Yujiu Yang, and Wanli Ouyang. 2025a. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*.

Runtao Liu, Haoyu Wu, Zheng Ziqiang, Chen Wei, Yingqing He, Renjie Pi, and Qifeng Chen. 2024. Videodpo: Omni-preference alignment for video diffusion generation. *arXiv preprint arXiv:2412.14167*.

Xialei Liu, Joost van de Weijer, and Andrew D. Bagdanov. 2017. Rankiqa: Learning from rankings for no-reference image quality assessment. *arXiv preprint arXiv:1707.08347*.

Yiming Liu, Jue Wang, Sunghyun Cho, Adam Finkelstein, and Szymon Rusinkiewicz. 2013. A no-reference metric for evaluating the quality of motion deblurring. *ACM Trans. Graph.*, 32(6).

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. 2025b. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

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

Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. 2013. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, and C.-C. Jay Kuo. 2015. Image database tid2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30:57–77.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Zeina Sinno and Alan Conrad Bovik. 2019. Large-scale study of perceptual video quality. *IEEE Transactions on Image Processing*, 28(2):612–627.

Wei Sun, Wen Wen, Xiongkuo Min, Long Lan, Guangtao Zhai, and Kede Ma. 2024. Analysis of video quality datasets via design of minimalistic video quality models. *2307.13981*.

Xinpeng Ti, Wentao Ye, Zhifang Zhang, Junbo Zhao, Chang Yao, Lei Feng, and Haobo Wang. 2025. Towards reverse engineering of language models: A survey. In *Findings of the Association for Computational Linguistics: EMNLP 2025*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Yibin Wang, Zhimin Li, Yuhang Zang, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. 2025a. Unified multimodal chain-of-thought reward model through reinforcement fine-tuning. *arXiv preprint arXiv:2505.03318*.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. 2025b. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*.

Yilin Wang, Sasi Inguva, and Balu Adsumilli. 2019. Youtube ugc dataset for video compression research. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, page 1–5. IEEE.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612.

Haoran Wei, Yaofeng Sun, and Yukun Li. 2025. Deepseek-ocr: Contexts optical compression. *2510.18234*.

Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. 2022. Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling. *arXiv preprint arXiv:2207.02595*.

Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. 2023a. Towards explainable in-the-wild video quality assessment: A database and a language-prompted approach. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 1045–1054. ACM.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. 2023b. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. 2023c. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*.

Tianhe Wu, Jian Zou, Jie Liang, Lei Zhang, and Kede Ma. 2025a. Visualquality-r1: Reasoning-induced image quality assessment via reinforcement learning to rank. *arXiv preprint arXiv:2505.14460*.

Zongyu Wu, Yuwei Niu, Hongcheng Gao, Minhua Lin, Zhiwei Zhang, Zhifang Zhang, Qi Shi, Yilong Wang, Sike Fu, Junjie Xu, and 1 others. 2025b. Lanp: Rethinking the impact of language priors in large vision-language models. *arXiv preprint arXiv:2502.12359*.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. 2022. Maniqa: Multi-dimension attention network for no-reference image quality assessment. *arXiv preprint arXiv:2204.08958*.

Zhenqiang Ying, Maniratnam Mandal, Deepti Ghadiyaram, and Alan Bovik. 2021. Patch-vq: 'patching up' the video quality problem. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 14014–14024. IEEE.

Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. 2019. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. *arXiv preprint arXiv:1912.10088*.

Zhiyuan You, Xin Cai, Jinjin Gu, Tianfan Xue, and Chao Dong. 2025. Teaching large language models to regress accurate image quality scores using score distribution. *arXiv preprint arXiv:2501.11561*.

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, Xiaocheng Feng, Jun Song, Bo Zheng, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2024. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. 2025a. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*.

Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. 2020. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1):36–47.

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

Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. 2023. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. *arXiv preprint arXiv:2303.14968*.

Xuanyu Zhang, Weiqi Li, Shijie Zhao, Junlin Li, Li Zhang, and Jian Zhang. 2025b. Vq-insight: Teaching vlms for ai-generated video quality understanding via progressive visual reinforcement learning. *arXiv preprint arXiv:2506.18564*.

Zhifang Zhang, Shuo He, Haobo Wang, Bingquan Shen, and Lei Feng. 2025c. Defending multimodal backdoored models by repulsive visual prompt tuning. *NeurIPS*.

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

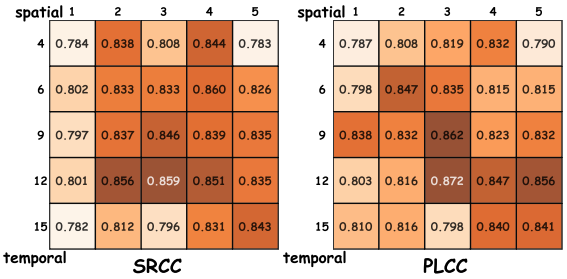


Figure 7: Performance on four synthetic VQA datasets (100 videos per dataset), peaking at temporal-flow images $M = 12$ and spatial-flow images $N = 3$ under hyper-parameter scanning.

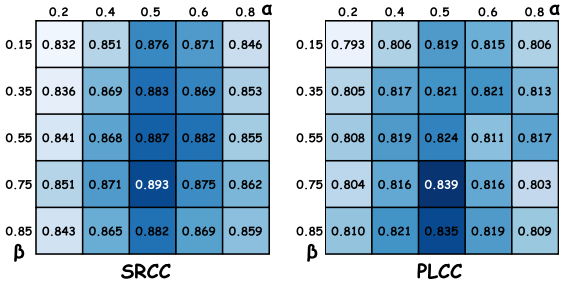


Figure 8: Performance on 10 synthetic IQA datasets (100 images per dataset) under hyper-parameter scanning of α and β.

A Hyperparameter scanning results

For **VQA** task, a **hyperparameter sweep** over M and N in Table 7) shows that performance increases with more global-temporal frames, peaking at $M = 12$ due to downsampling trade-offs. In contrast, once multiple local-spatial frames are jointly aggregated for reasoning ($N > 2$), performance stabilizes, highlighting the robustness of PreResIQA-R1 as the base for local evidence integration.

For **IQA** task, we conduct a **hyperparameter sweep** over α and β in Table 8. Both SRCC and PLCC remain stable across a broad range of values, indicating low sensitivity to hyperparameter choices, with peak performance achieved at $\alpha = 0.5$ and $\beta = 0.75$. Beyond scalar metrics, the error distributions in Fig. 5 show that PreResQ-R1 produces tighter and more centered predictions, offering stronger and more reliable evidence of its superiority in quality score estimation.

B Ablation experiment on reward

Tab. 6 further reveals the complementary, non-redundant nature of Response-dim and Preference-dim: gray-highlighted configurations integrating both dimensions outperform single-dimension setups, confirming joint optimization of response stability and preference alignment as the performance key. The optimal configuration (Res-Pre-dim in both stages) achieves SRCC 0.811±0.008 (+0.103) and PLCC 0.790±0.007 (+0.056), supporting our insight that response stabilization enables more effective preference learning.

Table 6: Ablation study on different design choices of Res-dim and Pre-dim across two training stages. Results are reported with SRCC/PLCC metrics, where values in parentheses denote improvements over the baseline.

Stage1	Stage2	SRCC	PLCC
Baseline	Baseline	0.708±0.051	0.734±0.043
w.Res-dim	w.Res-dim	0.721±0.006(+0.013)	0.749±0.006(+0.015)
w.Res-dim	w.Res-Pre-dim	0.789±0.008(+0.081)	0.783±0.007(+0.049)
w.Pre-dim	w.Pre-dim	0.754±0.042(+0.046)	0.762±0.037(+0.028)
w.Pre-dim	w.Res-Pre-dim	0.787±0.010(+0.079)	0.780±0.009(+0.046)
w.Res-Pre-dim	w.Res-Pre-dim	0.811±0.008(+0.103)	0.790±0.007(+0.056)
w.Res-Pre-dim	w.Res-dim	0.809±0.015(+0.101)	0.785±0.014(+0.051)

Quantitatively, Response-dim and Preference-dim contribute 37% and 50% to performance gains, respectively: Preference-dim enhances global human preference alignment, while Response-dim suppresses intra-sample variability, stabilizes optimization, and boosts generalization. Critical for MLLM-based QA under limited supervision (6K images, 28K videos), Response-dim ensures stable performance across 10 benchmarks. These results confirm the rationality of PRPO and the necessity of joint response-preference optimization to address trajectory instability and distribution mismatch overlooked by existing methods.

C Evaluation of Reasoning Quality

As shown in Table 8, PreRes-IQA and PreRes-VQA consistently outperform their respective baselines across all four GPT-5 mini reasoning dimensions: Accuracy, Reasoning, Completeness, and Confidence. This indicates that the model generates structured and coherent CoT outputs that capture meaningful perceptual cues beyond simple score prediction.

D Temporal-Spatial Flow Effectiveness

In addition, the ablation results in Table 9 show substantial improvements in SRCC and PLCC when incorporating the Preference-Response Disentangled Policy Optimization and Temporal-Spatial Flow modules. This confirms that TSF effectively aggregates global and local visual evidence for video quality assessment, while PRPO stabilizes intra-sample responses and aligns inter-sample preferences.

Table 7: Performance of PreResQ-R1 on KonIQ datasets across SRCC and PLCC. **Red**, **green**, and **blue** denote handcrafted, discriminative deep-learning, and MLLM-based methods. **Bold** and underline denote the best and second best results. All results are obtained by averaging over ten independent runs with different random seeds.

Method	SPAQ		BID		CLIVE		Liu13		SRIQA		TID13		KonIQ		KADID		PIPAL		AQIOA		Avg	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
BRISQUE	0.406	0.490	0.329	0.337	0.256	0.267	0.352	0.347	0.187	0.149	0.531	0.550	0.226	0.225	0.356	0.429	0.232	0.267	0.497	0.541	0.337	0.360
NIQE	0.664	0.679	0.394	0.405	0.309	0.329	0.309	0.317	0.215	0.190	0.509	0.521	0.533	0.530	0.405	0.468	0.161	0.195	0.533	0.560	0.403	0.419
MUSIQ	0.863	0.868	0.516	0.513	0.287	0.312	0.347	0.360	0.254	0.227	0.627	0.634	0.929	0.924	0.556	0.575	0.431	0.431	0.630	0.720	0.544	0.557
ManIQA	0.768	0.864	0.492	0.510	0.291	0.305	0.494	0.485	0.234	0.225	0.681	0.673	0.758	0.768	0.499	0.654	0.457	0.419	0.723	0.685	0.540	0.559
CLIP-IQA+	0.866	0.758	0.527	0.536	0.356	0.391	0.515	0.524	0.459	0.439	0.643	0.652	0.864	0.866	0.653	0.465	0.427	0.452	0.736	0.636	0.605	0.572
Qwen2.5-VL	0.869	0.872	0.711	0.725	0.604	0.635	0.742	0.738	0.441	0.418	0.702	0.682	0.899	0.889	0.675	0.693	0.384	0.375	0.752	0.785	0.677	0.681
DeQA	0.896	0.895	0.716	0.714	0.616	0.662	0.761	0.749	0.510	0.504	0.659	0.661	0.941	0.953	0.687	0.694	0.478	0.472	0.815	0.814	0.708	0.712
C2Score	0.860	0.867	0.725	0.729	0.627	0.657	0.775	0.765	0.431	0.409	0.726	0.751	0.881	0.875	0.687	0.694	0.342	0.345	0.729	0.809	0.678	0.690
Q-Align	0.887	0.886	0.707	0.719	0.694	0.713	0.805	0.791	0.479	0.428	0.754	0.769	0.940	0.941	0.684	0.674	0.419	0.403	0.735	0.772	0.710	0.710
Q-Insight	0.905	**0.907**	0.755	0.759	0.715	0.751	**0.815**	0.806	0.561	**0.550**	0.729	0.732	0.916	0.933	0.736	0.742	0.486	0.474	0.764	0.811	0.738	0.747
VisualQuality-R1	0.892	0.891	0.743	0.751	0.749	0.776	0.784	0.785	0.554	0.531	0.769	0.775	0.915	0.924	0.711	0.712	0.438	0.441	0.867	0.822	0.742	0.741
PreResQ-R1 (Ours)	**0.919**	0.904	**0.761**	**0.765**	**0.762**	**0.792**	0.810	**0.808**	**0.579**	0.541	**0.774**	**0.781**	**0.951**	**0.959**	**0.741**	**0.750**	**0.491**	**0.482**	**0.882**	**0.835**	**0.767**	**0.762**

Table 8: Quantitative evaluation on the GPT-5 mini benchmark over four reasoning dimensions: Accuracy, Reasoning, Completeness, and Confidence. Higher scores denote better performance.

Model	Accuracy	Reason	Completeness	Confidence
VisualQuality-R1	3.94	4.69	4.71	4.66
PreRes-IQA	**4.65**	**5.21**	**5.85**	**5.15**
VQAThinker	3.62	3.94	3.82	4.15
PreRes-VQA	**3.94**	**4.71**	**4.62**	**4.59**

Table 9: Ablation study on the effectiveness of proposed Preference-Response disentangled Policy Optimization (PRPO) and Temporal-Spatial Flow (TSF) modules. Values in parentheses indicate the improvements over the baseline.

Model	SRCC	PLCC
Baseline	0.767	0.621
w. PRPO	0.819(+53%)	0.793(+68%)
w. TSF	0.865(+47%)	0.872(+32%)

E Performance on KonIQ dataset

We further assess cross-dataset generalization to evaluate the robustness and transferability of PreResQ-R1. As reported in Table 7, when trained solely on KonIQ, PreResQ-R1 achieves an average SRCC of 0.758 and PLCC of 0.743 across four distinct IQA datasets, corresponding to improvements of 12.70% and 7.09% over the Qwen2.5-VL baseline. These results indicate that the preference-response disentangled reinforcement learning framework not only stabilizes intra-sample reasoning but also facilitates the learning of perceptual representations that generalize effectively across diverse visual domains. The substantial gains in cross-dataset performance highlight the model's ability to disentangle and robustly encode attribute-level quality cues, suggesting that the learned representations capture fundamental perceptual principles rather than overfitting to dataset-specific score distributions.

Table 10: Fine-grained CoT performance on SRCC metric in IQA task.

Task	Saturation	Granularity	Sharpness	Foreground	Background	SRCC
IQA	0.613	0.431	0.470	0.564	0.264	0.811

Table 11: Global-local CoT performance on SRCC metric in VQA task.

Task	Global	Local	SRCC
VQA	0.619	0.284	0.865

F Perceptual-Level Performance Analysis

The results in Tables 10 and 11 reveal nuanced insights into PreResQ-R1's handling of perceptual dimensions across IQA and VQA tasks. For IQA, while the overall SRCC reaches 0.811, the contribution of individual attributes varies considerably: saturation (0.613) and foreground quality (0.564) align strongly with human judgments, whereas background fidelity (0.264) and granularity (0.431) show weaker correlations. This divergence indicates that PreResQ-R1 captures heterogeneous, attribute-specific visual cues that are not trivially aggregated into the final score, reflecting the model's ability to maintain discriminative intermediate representations and disentangle fine-grained perceptual factors from global assessments. In VQA, the global-temporal context exhibits higher predictive strength (SRCC 0.619) than local-spatial cues (0.284), yet their combination reaches 0.865, highlighting the model's capacity to integrate complementary evidence across temporal and spatial scales. These findings suggest that robust video quality assessment benefits not from over-reliance on either temporal coherence or frame-level detail, but from structured aggregation that preserves the informational distinctiveness of each dimension. It demonstrates that PRPO not only stabilizes global predictions but also cultivates interpretable, multi-dimensional perceptual representations.

G Further methodological insights

Spatiotemporal aggregation in VQA: global-temporal and local-spatial decoupling. Inspired by DeepSeek-OCR (Wei et al., 2025), which compresses textual sequences into compact image tokens, we extend this principle to video quality assessment by decoupling temporal and spatial evidence. The global-temporal branch captures overall motion consistency and smoothness by compressing multiple frames into a single multiframe representation, enabling efficient reasoning over temporal coherence. The local-spatial branch evaluates perceptual quality within individual frames, sampling three key frames to balance accuracy and computational cost. Unlike prior methods such as MinimalisticVQA (Sun et al., 2024) and VQA-Thinker (Cao et al., 2025a), which rely on downsampling or reinforcement for temporal consistency, our approach explicitly disentangles temporal dynamics from spatial fidelity. Pretraining on image-level quality provides strong initialization for video-level assessment, demonstrating that decoupled aggregation combined with cross-modal transfer yields robust and interpretable video quality evaluation.

Why we divide visual assessment into fine-grained details. Human perception of visual quality is intrinsically multi-dimensional, encompassing interrelated factors such as texture fidelity, color saturation, contrast, and semantic clarity. Conventional IQA models that regress a single holistic score fail to capture these heterogeneous perceptual cues, leading to entangled representations and unstable reasoning. By decomposing assessment into five fine-grained perceptual dimensions (Saturation, Granularity, Sharpness, Foreground, and Background) PreResQ-R1 explicitly structures the visual reasoning process. This decomposition enforces a dimension-aware reasoning pathway that aligns linguistic explanations with visual evidence, thereby promoting fine-grained perceptual reasoning. The structured formulation further stabilizes optimization through dimension-wise regularization (e.g., standard deviation penalty) and reward balancing, ultimately enhancing interpretability, robustness, and perceptual fidelity.

Why we introduce a preference-response disentangled policy optimization definition. Conventional reward formulations in RL-based IQA reduce perceptual alignment to either scalar regression or ordinal ranking, neglecting the intertwined nature of intra-sample reasoning stability and inter-sample perceptual consistency that underpin human quality judgment. The proposed Preference-Response Disentangled Policy Optimization establishes a dual-axis reward geometry that decomposes optimization into response coherence and preference alignment. The former regularizes reasoning trajectories within each perceptual context, while the latter constrains global quality ordering across heterogeneous samples. This disentangled formulation transforms reward learning from distribution-specific scalar fitting into a structurally grounded perceptual reasoning process, preventing spurious reward exploitation and overfitting to dataset priors. By explicitly balancing local response fidelity and global perceptual regression, PRPO enforces reward discrimination, enhances policy robustness against distributional shift.

Why we train only on KADID-10K dataset, and evaluate PreResQ-R1 on ten datasets. Training solely on KADID-10K deliberately targets preference alignment rather than re-teaching visual perception: we assume the underlying MLLM already possesses strong visual understanding and reasoning priors, and **our goal is to use minimal RL fine-tuning to correct its deviations from human aesthetic preference**. This low-cost RL calibration nudges the model's pre-trained perceptual priors toward human MOS with limited supervision. Evaluation across ten heterogeneous benchmarks that cover both synthetic and authentic distortions constitutes a rigorous out-of-distribution test. The consistently high SRCC and PLCC values illustrate that our disentangled PRPO framework effectively mitigates human-model preference bias while retaining the inherent generalization ability of the MLLM. In short, this work emphasizes efficient alignment (minimal RL intervention) over wholesale perceptual re-learning, offering a scalable path for human-aligned aesthetic assessment.

Why we voluntarily introduce std penalty and response balance reward to obtain Exploration-to-Stability fine-tuning process. We introduce a standard deviation penalty at the fine-grained answer level in first exploration stage to accelerate the model's sensitivity to differences across granular aspects, enabling it to accurately predict each fine-grained component. The response balance reward promotes consistent and stable generations, which not only enhances overall training stability but also supports reliable preference ranking and score-based rewards. These mechanisms ensure

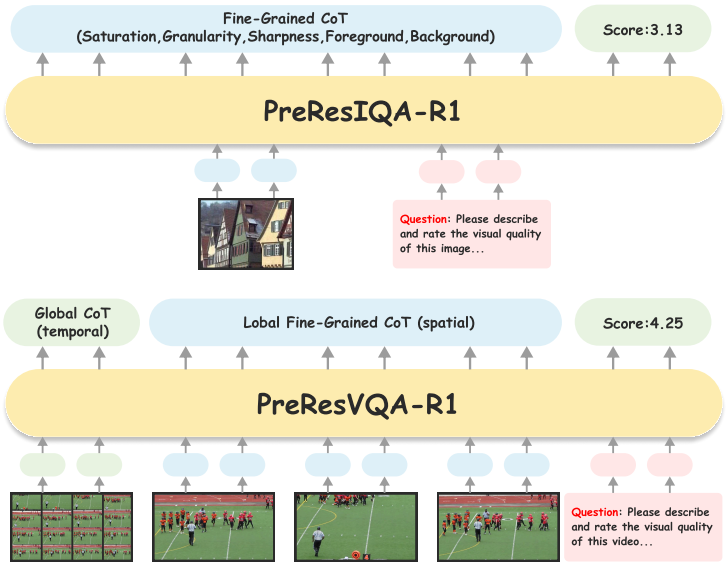


Figure 9: The input and output data flow in PreResIQA-R1 and PreResVQA-R1.

that the model learns nuanced distinctions while maintaining coherent and calibratable outputs.

Why we introduce the preference triplet ranking score with the introduction of pairwise ranking and score reward. While pairwise comparisons capture local perceptual consistency, they are insufficient to enforce global transitivity across multiple samples. The preference triplet ranking-score reward introduces a challenging and fine-grained relational constraint by modeling third-order transitivity, guiding the policy to refine subtle perceptual distinctions and construct globally coherent hierarchies. By integrating pairwise and triplet consistency, we form a synergistic reward landscape in which pairwise terms anchor local human alignment, while triplet terms eliminate cyclic inconsistencies and enhance global ranking stability.

Multi-prompts in exploration-fine-tuning stage: why it works. In the early exploration phase, fixed prompts tend to restrict reasoning diversity and confine the model's attention to narrow perceptual patterns. Introducing multiple semantically varied prompts (e.g., focusing on "texture clarity," "color naturalness," or "composition balance") stimulates broader CoT trajectories and exposes the model to different linguistic formulations of quality reasoning. This multi-prompt strategy acts as reward-level data augmentation, allowing the policy to explore a richer reasoning space and uncover diverse perceptual cues before convergence. Subsequent stability refinement distills this diversity into a calibrated policy, improving both reasoning coherence and generalization to unseen distortions.

Two-stage exploration-stability fine-tuning's design considerations. The "exploration-stability" fine-tuning paradigm shares substantial similarities with DeepSeek-R1 (DeepSeek-AI, 2025), which combines multi-prompt exploration with stability-driven reinforcement learning refinement. In general, early exploration allows the model to probe diverse reasoning trajectories and perceptual cues, while the stability phase consolidates these experiences into a consistent policy. However, without clear evidence of modifications or novel mechanisms beyond prior two-stage frameworks, it is difficult to attribute unique methodological advances to this design choice. Careful justification and ablation are therefore needed to substantiate the novelty and effectiveness of such two-stage fine-tuning.

H Prompt templates and qualitative cases

We provide a detailed overview of the overall algorithmic workflow in Algorithm 1, and present the complete prompt templates used for both IQA and VQA tasks in Table 12. To further complement our main results, we include extensive qualitative analyses and ablation studies in the supplementary figures. Specifically, Fig. 11, Fig. 13, and Fig. 12 illustrate additional comparison cases with state-of-the-art methods, as well as preference-response ablation analyses and step-wise training behaviors for the IQA task. Similarly, Fig. 14, Fig. 15, and Fig. 16 present corresponding results for the VQA task, offering a more comprehensive understanding of our model's reasoning ability, robustness, and cross-domain generalization.

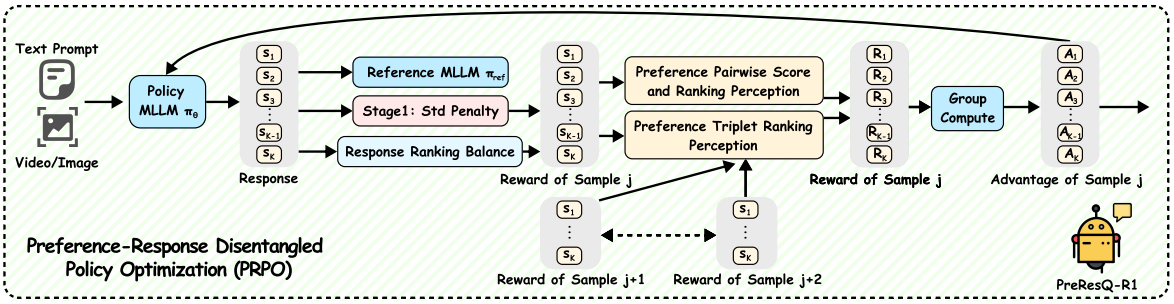


Figure 10: Pipeline of the Preference-Response Disentangled Policy Optimization (PRPO), which introduced response-based ranking reward, preference-based pairwise score-ranking reward, and preference-based triplet ranking reward to optimize group policy learning.

I Reinforcement learning optimization

We adopt a regularized policy optimization framework to fine-tune the MLLM using the proposed reward. For each input sample j, we generate K stochastic trajectories and collect their total rewards into a vector:

$$\mathbf{r}(s_j) = [R_{\text{total}}^{j,1}, \ldots, R_{\text{total}}^{j,K}]^\top. \tag{11}$$

To reduce variance across stochastic generations, we compute a normalized relative advantage:

$$a(s_j^{(i)}) = \frac{R_{\text{total}}^{j,i} - \mu(\mathbf{r}(s_j))}{\sigma(\mathbf{r}(s_j))}, \tag{12}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ denote the mean and standard deviation over the K trajectories of sample j. We optimize the policy π_θ using a clipped surrogate objective with KL regularization to a fixed reference policy π_{ref}, following standard practice:

$$\begin{aligned}\mathcal{L}(\theta) = \frac{1}{BK}\sum_{j=1}^{B}\sum_{i=1}^{K}\min\Bigg[&\frac{\pi_\theta(o_i|c,s_j)}{\pi_\theta^{\text{old}}(o_i|c,s_j)}a(s_j^{(i)}),\\ &\text{clip}\left(\frac{\pi_\theta(o_i|c,s_j)}{\pi_\theta^{\text{old}}(o_i|c,s_j)}, 1-\epsilon, 1+\epsilon\right)a(s_j^{(i)})\Bigg] - \lambda D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})\,.\end{aligned} \tag{13}$$

where ϵ is the clipping threshold and λ controls the strength of KL regularization. The KL divergence is approximated as:

$$D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}}) \approx \pi_{\text{ref}}\frac{\pi_\theta}{\pi_{\text{ref}}} - \log\frac{\pi_\theta}{\pi_{\text{ref}}} - 1, \tag{14}$$

which constrains the updated policy to remain close to the pretrained MLLM. Finally, policy optimization follows the standard policy gradient formulation:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\Big[R_{\text{total}}\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(a_t|s_t)\Big], \tag{15}$$

where a_t denotes the generated token at step t and $s_t = (\mathcal{P}, \mathcal{I}, a_{<t})$ is the corresponding state.

J Ethical considerations

This work aims to improve the reliability of visual quality assessment in multimodal systems, but several risks remain. First, QA models reflect subjective human preferences, and biases in training data may propagate into systematic disparities across content or demographic groups. Second, increased scoring stability may obscure underlying uncertainty, potentially leading to overconfident judgments in sensitive applications. Third, large multimodal models incur nontrivial computational and environmental costs. We stress that PreResQ-R1 is intended as a research framework rather than a definitive metric, and requires careful validation before deployment in real-world settings.

Table 12: Fine-grained structured text prompts for IQA and VQA tasks.

IQA Prompt 1
You are doing an image quality assessment. Provide five ratings in the following order: saturation, granularity, sharpness, foreground, and background. Please rate 5 scores. Each rating should be a float between 1 and 5, rounded to two decimals (1 = very poor, 5 = excellent).
IQA Prompt 2
You are doing an image quality assessment. For this image quality assessment task, please evaluate along five aspects: color saturation, texture granularity, clarity or sharpness, foreground quality, and background quality. Please rate 5 scores each row. Give each aspect a score between 1.00 and 5.00 (two decimals, where 1 = very poor and 5 = excellent).
IQA Prompt 3
You are doing an image quality assessment. Please assess image quality by providing five ratings in sequence: rate the saturation, rate the granularity, rate the sharpness, rate the foreground quality, and rate the background quality. Please rate 5 scores. All answers must be floats from 1.00 to 5.00, rounded to two decimals.
IQA Prompt 4
You are doing an image quality assessment. Provide ratings for five dimensions: saturation to reflect color vividness, granularity to capture texture smoothness, sharpness to evaluate detail clarity, foreground quality to assess object visibility, and background quality to measure scene consistency. Please rate 5 scores each row. Each score must be a float between 1.00 and 5.00, where 1 means poor and 5 means excellent.
IQA Prompt 5
You are doing an image quality assessment. Please give an image quality evaluation with five scores in the following order: saturation, granularity, sharpness, foreground quality, and background quality. Please rate 5 scores each row. Each score should be a float between 1 and 5, rounded to two decimals, with 1 indicating very poor quality and 5 indicating excellent quality.
IQA Template
You are doing the image quality assessment task. Please describe and rate the picture in five parts: saturation rating; granularity rating; sharpness rating; foreground rating; and background rating. All ratings should be floats between 1 and 5, rounded to two decimals, where 1 represents very poor quality and 5 represents excellent quality. <think> ...Reasoning CoT... < /think> <answer>Score1; Score2; Score3; Score4; Score5< /answer>
VQA Template
You are doing the video quality assessment task. Please compare the internal packed video frames in the first picture and describe the global multiframe pictures in temporal deminsion and global scene. Then compare and describe the rest local single frame pictures in five parts: saturation rating; granularity rating; sharpness rating; foreground rating; and background rating. Finally, please rate two scores in the global multiframe and local single frame dimension. All ratings should be floats between 1 and 5, where 1 represents very poor quality and 5 represents excellent quality. <think> ...Reasoning CoT... < /think> <answer>Score1; Score2< /answer>

Algorithm 1 PreResQ-R1: Two-stage fine-grained rank-and-score reinforcement learning for image quality assessment via preference-response disentangled policy optimization.

Input: Training set $\mathcal{D} = \{(\mathcal{I}_j, \mathcal{P}_j, \mathbf{G}_j)\}_{j=1}^{B}$, pretrained policy π_θ, reference policy π_{ref}, generation number K, batch size B, total steps T, reward weights $\alpha, \beta_{\text{err}}, \beta_{\text{rank}}$, and standard deviation penalty parameters $\lambda_{\text{std}}, \Omega_{\text{std}}$.

Output: Updated policy π_θ.

Stage 1: Multi-prompt Exploration with Std Penalty
for each image or video sample $\mathcal{I}_i$ in dataset **do**
  Randomly select a prompt $\mathcal{P}_x$ from predefined set $\{\mathrm{P_I}, ..., \mathrm{P_V}\}$.
  Generate K responses $\tau_j^{(1:K)} \sim \pi_\theta(\cdot|\mathcal{P}_x, \mathcal{I}_j)$.
  Extract dimension-wise scores $\mathbf{s}_j^{(i)} = [s_j^1, ..., s_j^5]$.
  Compute std penalty $R_{\text{std}}(\mathbf{s}_j^{(i)})$ as Eq. (16).
  Compute preference-response disentangled rewards $R_{\text{for}}, R_{\text{loc}}^{j,i}, R_{\text{pair}}^{j,i}, R_{\text{tri}}^{j,i}, R_{\text{std}}$ (Eq. (2-9)).
  Form total reward $R_j^{(i)}$ as Eq. (10-11).
  Update π_θ via PRPO objective (Eq. (12-15)).
end for
▷ *Exploration RL2RS fine-tuning with PRPO optimization.*
Stage 2: Single-prompt Stability Refinement
for each image or video sample $\mathcal{I}_i$ in dataset **do**
  Use single standard prompt $\mathcal{P}$ for generation.
  Generate K responses $\tau_j^{(1:K)} \sim \pi_\theta(\cdot|\mathcal{P}, \mathcal{I}_j)$.
  Extract predicted scores $\mathbf{s}_j^{(i)}$.
  Compute rewards $R_{\text{for}}, R_{\text{loc}}^{i,k}, R_{\text{pair}}^{i,k}, R_{\text{tri}}^{i,k}$ (Eq. (2-9)).
  Obtain final reward $R_j^{(i)}$ and normalize advantages $a_j^{(i)}$ (Eq. (10-11)).
  Update π_θ via PRPO objective (Eq. (12-15)).
end for
▷ *Stability RL2RS fine-tuning with PRPO optimization.*
return Optimized policy π_θ.



Figure 11: Qualitative comparison cases of image quality assessment results between PreResQ-R1 and state-of-the-art reinforcement learning based reasoning MLLM. Given an image, and a text prompt, our model performs quality assessment across saturation, granularity, sharpness, foreground, and background through CoT reasoning.

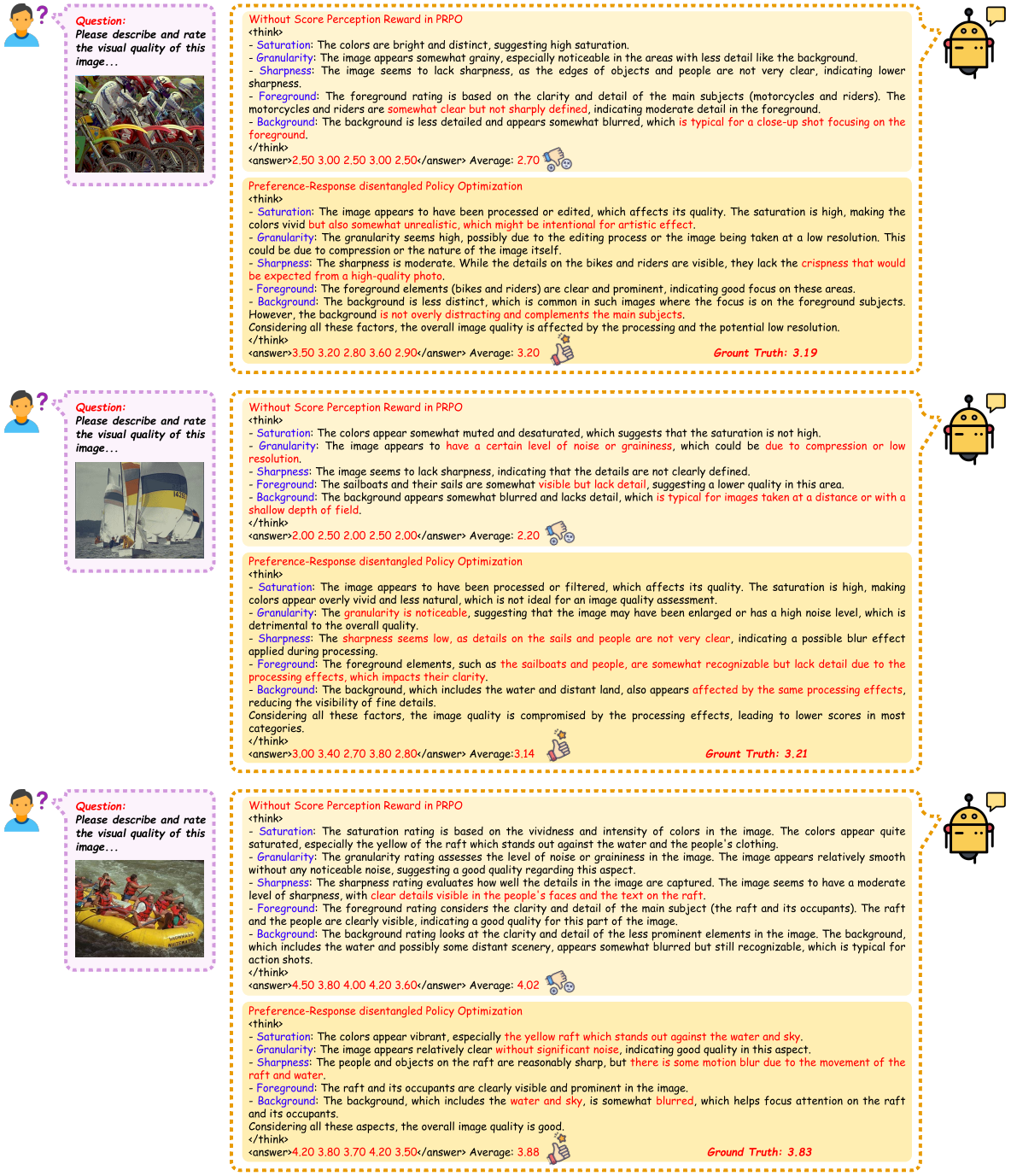


Figure 12: Qualitative ablation cases of PreResQ-R1's image quality assessment results with/without Score Perception Reward in Preference-Response disentangled Policy Optimization.

Question:
Please describe and rate the visual quality of this image...



<think>
The image appears to have some noise and lacks sharpness, which affects the overall clarity. The colors seem somewhat muted, and there is a noticeable graininess that detracts from the image's quality. The main subject (the boat) is recognizable but not sharply defined. The background is less distinct due to the same issues of noise and lack of sharpness.
</think> <answer> 2.80 2.70 2.60 2.90 2.30 </answer> Average: 2.66 Training Step: 100

<think>
The image appears to have a somewhat muted color palette, which suggests lower saturation. The image seems to have a grainy texture, indicating higher granularity. The details of the boat and the water are not very clear, suggesting lower sharpness. The foreground is the boat itself, which is somewhat clear but lacks detail due to the graininess. The background includes the water and some distant land, which are also not very sharp and lack detail.
</think> <answer> 1.00 2.00 2.00 3.50 3.00 </answer> Average: 2.30 Training Step: 200

<think>
To assess the image quality, I'll evaluate it based on the following criteria:
- Saturation: The colors appear somewhat muted and not very vibrant.
- Granularity: There is noticeable noise or graininess in the image.
- Sharpness: The details of the boat and the surrounding environment are not very clear.
- Foreground: The boat is the main subject but lacks detail due to the image quality.
- Background: The distant land and sky are also not very sharp.
Based on these observations, I will assign ratings accordingly.
</think> <answer> 3.67 3.45 3.20 3.89 3.10 </answer> Average: 3.46 Training Step: 400

<think>
The saturation rating considers the vividness and intensity of colors in the image. The colors appear somewhat muted, which is common in older photographs or images taken under certain lighting conditions.
The granularity rating assesses the level of noise or graininess in the image. The image appears slightly grainy, which could be due to the age of the photograph or the quality of the camera used at the time.
The sharpness rating evaluates how well-defined the edges and details in the image are. The image lacks sharpness, which might be due to the age of the photograph or the focus settings during capture.
The foreground rating considers the clarity and prominence of the main subject. The boat is the main subject and is somewhat clear, but the lack of sharpness affects its prominence.
The background rating evaluates the clarity and detail in the less prominent areas of the image. The background, including the water and distant land, appears somewhat blurry and lacks detail.
Considering all these factors, here are the ratings:
</think><answer>3.00 3.00 2.80 3.50 3.20 </answer> Average: 3.10 Training Step: 800

<think>
- Saturation: The colors appear somewhat muted, suggesting that the saturation is not very high.
- Granularity: There appears to be some noise or graininess, which affects the overall clarity.
- Sharpness: The image lacks sharpness, indicating that details are not clearly defined.
- Foreground: The boat is the main subject, but it appears somewhat faded and lacks detail, which affects its prominence.
- Background: The background includes water and trees, both of which lack sharpness and detail, contributing to the overall lower quality.
Based on these considerations:
</think> <answer> 3.00 3.20 3.10 3.40 3.60 </answer> Average: 3.26 Training Step: 1600 ***Ground Truth: 3.25***

Question:
Please describe and rate the visual quality of this image...



<think>
The image appears clear with good detail visible in the lighthouse, surrounding structures, and the sky. The colors are vibrant, suggesting good saturation and the image shows a good depth of field, which indicates a decent sharpness. The texture of the rocks and the lighthouse structure is well-captured, suggesting good granularity. The overall composition and foreground (lighthouse and buildings) are well-defined, giving it a higher rating for the foreground. The background consists of the sky, which is relatively uniform but complements the foreground nicely, bringing up the background rating.
</think> <answer> 4.50 4.30 4.20 4.40 4.10 </answer> Average: 4.34 Training Step: 100

<think>
For the saturation rating, the image appears to have a strong contrast which indicates high saturation. The colors are vivid and the sky and ocean have clear differentiation from the buildings.
For the granularity rating, the image does not show any visible noise or grain, which suggests a smooth and clean image quality.
For the sharpness rating, the details of the lighthouse, building, and surrounding rocks are clear and well-defined, indicating high sharpness.
For the foreground rating, the rocks in the foreground are in focus and provide depth to the image, which is positive.
For the background rating, the background elements, including the sky and water, are clear and do not detract from the main subject.
Overall, the image appears to be of high quality in terms of color, contrast, and clarity.
</think> <answer>4.00 4.00 4.50 4.25 4.50 </answer> Average: 4.25 Training Step: 200

<think>
The image quality appears to be generally good with no visible distortions or artifacts. Here's the analysis:
- Saturation: The colors are vivid and consistent, which is a good indication of high saturation.
- Granularity: The image seems to have a uniform texture without any visible noise or graininess, indicating good quality.
- Sharpness: The edges of the lighthouse, the house, and the surrounding details appear distinct and well-defined, suggesting high sharpness.
- Foreground: The rocky foreground is detailed and clearly visible.
- Background: The sky and the water in the distance are clear and expansive.
</think> <answer> 4.30 4.25 4.15 4.25 4.05 </answer> Average: 4.20 Training Step: 400

<think>
- Saturation: The saturation rating considers the vividness and intensity of colors in the image. The colors in this image appear natural and well-balanced without any overly saturated or muted tones.
- Granularity: The granularity rating assesses the level of noise or graininess in the image. The image appears clear with no significant noise, indicating good quality.
- Sharpness: The sharpness rating evaluates how well-defined and clear the edges and details are in the image. The image shows clear edges and details, especially in the lighthouse and the surrounding structures.
- Foreground: The foreground rating considers the clarity and prominence of the main subject. The lighthouse is the main subject and is clearly visible with good detail.
- Background: The background rating considers the clarity and detail of the less prominent elements. The sky and water in the background are clear and contribute to the overall composition without distracting from the main subject.
Considering all these aspects, the image quality seems high.
</think><answer>4.25 4.00 4.25 4.50 4.00</answer> Average: 4.20 Training Step: 800

<think>
- Saturation: The colors appear natural and well-balanced, suggesting good saturation.
- Granularity: The image appears relatively smooth without any noticeable noise, indicating a high level of detail.
- Sharpness: The details in the lighthouse and the surrounding structures are clear, suggesting a high level of sharpness.
- Foreground: The lighthouse and the nearby buildings are in focus, with clear details visible, indicating a good foreground quality.
- Background: The sky and the distant water are clear but not as detailed as the foreground, which is typical for landscape photography.
</think><answer>4.20 4.00 4.30 4.10 4.20</answer> Average: 4.16 Training Step: 1600 ***Ground Truth: 4.16***

Figure 13: Qualitative comparison cases of PreResQ-R1's image quality assessment results across different training steps.

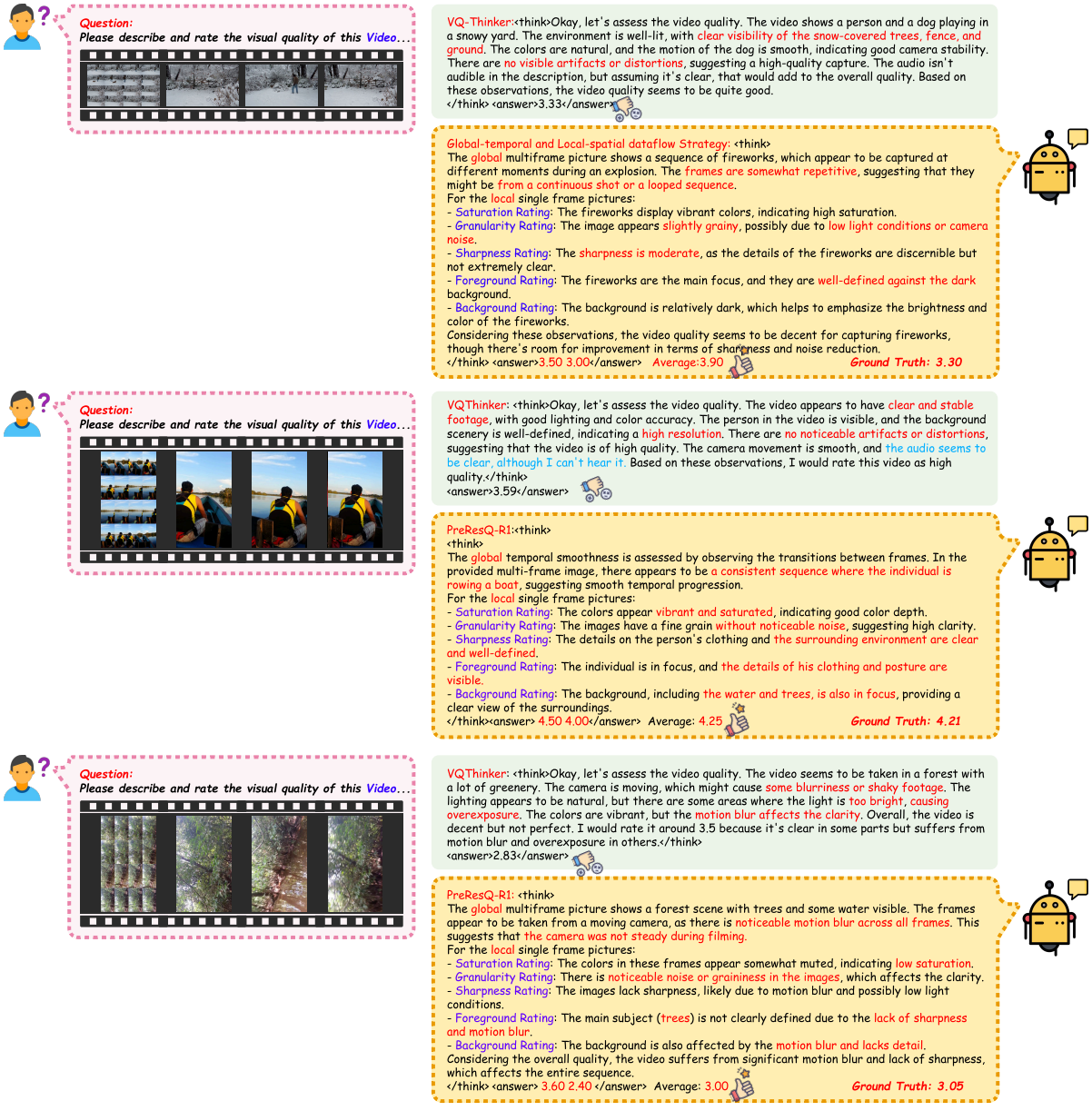


Figure 14: Qualitative comparison cases of video quality assessment results between PreResQ-R1 and state-of-the-art reinforcement learning based reasoning MLLM. Given an video sample, and a text prompt, our model performs quality assessment across saturation, granularity, sharpness, foreground, and background through CoT reasoning.

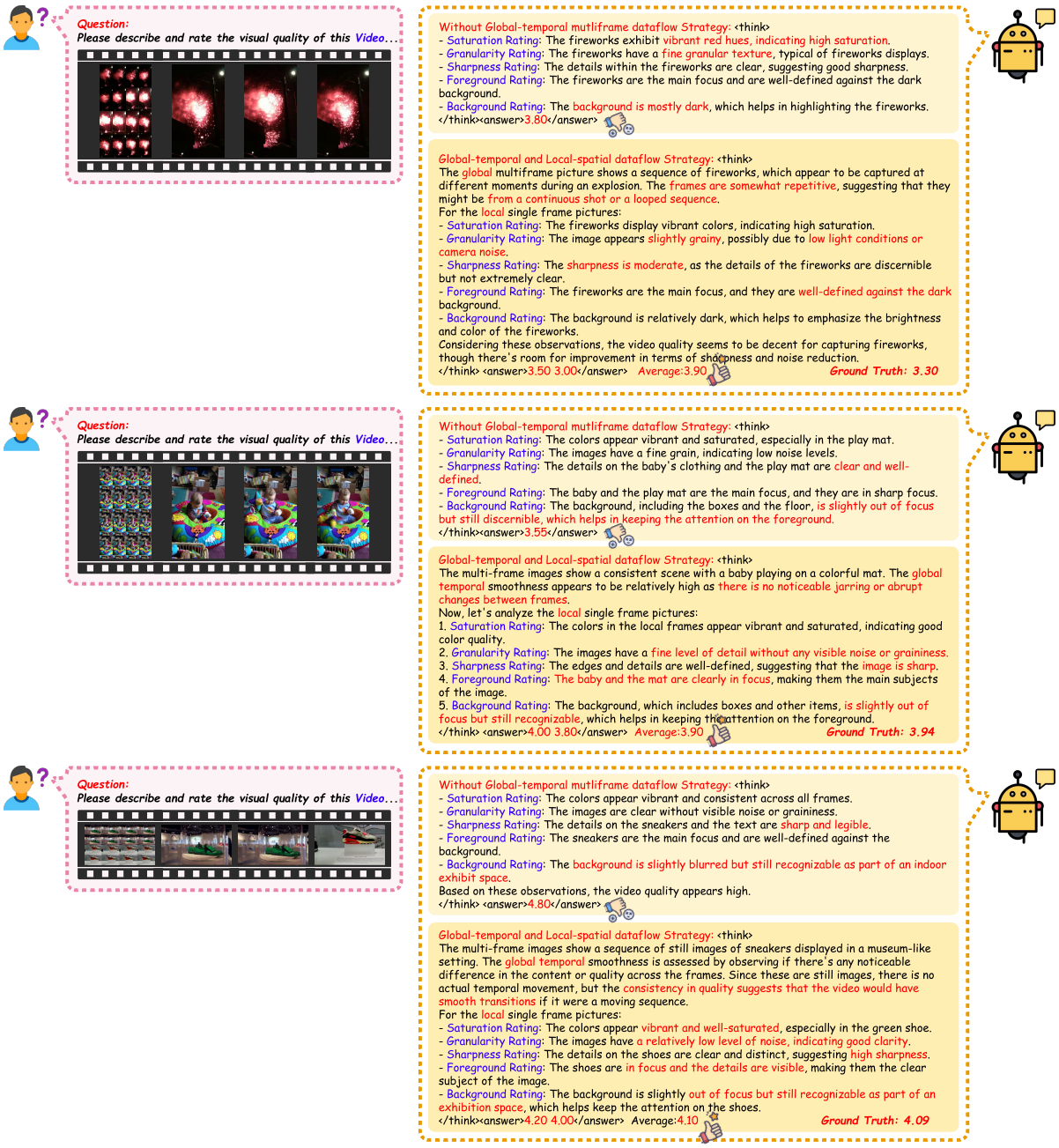


Figure 15: Qualitative ablation cases of PreResQ-R1's video quality assessment results with/without Global-Temporal dimension CoT in reasoning process.



Figure 16: Qualitative comparison cases of PreResQ-R1's video quality assessment results across different training steps.

K Compared Methods

1. **NIQE** (Mittal et al., 2013): a handcrafted no-reference IQA model that assesses image quality by measuring deviations from natural scene statistics, without relying on any human-rated training data or prior knowledge of distortions.

2. **STEM** (Kancharla and Channappayya, 2022): a training-free method for assessing the quality of user-generated videos. Inspired by the human visual system's perceptual straightening hypothesis, STEM measures how distortions increase the curvature of a video's perceptual trajectory.

3. **BRISQUE** (Mittal et al., 2012): a handcrafted NR-IQA model that evaluates image quality by measuring deviations from natural scene statistics of locally normalized luminance in the spatial domain, without requiring distortion-specific features or reference images.

4. **UNIQUE** (Zhang et al., 2021): a deep learning-based BIQA model that learns to assess image quality through pairwise comparison and fidelity loss, enabling unified training across both synthetic and real-world distortions for robust generalization.

5. **MUSIQ** (Ke et al., 2021): a deep learning-based IQA model that leverages a multi-scale Transformer architecture to assess image quality directly from images of arbitrary resolutions, capturing quality information at multiple granularities without the need for resizing or cropping.

6. **MANIQA** (Yang et al., 2022): a deep learning-based NR-IQA model that employs a multi-dimensional attention architecture combining channel- and spatial-attention mechanisms (via Transposed Attention Block and Scale Swin Transformer Block) to jointly capture global and local visual dependencies, achieving superior performance—especially on GAN-distorted images.

7. **LIQE** (Zhang et al., 2023): a MLMM-based BIQA method that leverages a multitask learning framework for joint learning of image quality assessment, scene classification, and distortion identification. By aligning visual and textual embeddings via cosine similarity, the model enables automatic parameter sharing and loss weighting, leading to more robust and generalizable quality prediction.

8. **FAST-VQA** (Wu et al., 2022): is an efficient deep video quality assessment framework that introduces a Grid Mini-patch Sampling strategy to retain both local and global quality information while drastically reducing computational cost. By processing these quality-preserving fragments through a specialized Fragment Attention Network, FAST-VQA enables effective end-to-end learning of video-quality-related representations.

9. **VQA^2-Scorer** (Jia et al., 2024): is a VQA model in the VQA2 series that leverages large multi-modal model architectures to predict quantitative quality scores for videos. It processes both spatial and temporal information by interleaving visual and motion tokens, enabling detailed perception of distortions and artifacts. Built upon the VQA2 Instruction Dataset, it is trained to capture nuanced quality differences across diverse video types, achieving strong performance in both traditional quality scoring and holistic visual quality understanding tasks.

10. **VQ-Insight** (Zhang et al., 2025b): is a reasoning-driven MLLM framework designed for evaluating the quality of AI-generated videos (AIGC). It introduces a progressive video quality learning paradigm: starting from image-level warm-up, advancing to temporal understanding, and finally achieving joint optimization with video generators. Moreover, it incorporates multi-dimensional scoring, preference comparison, and temporal modeling rewards to balance generalization and specialization. Extensive experiments show that VQ-Insight outperforms state-of-the-art methods across multiple evaluation settings, offering reliable and interpretable video quality assessment for AIGC applications.

11. **VQThinker** (Cao et al., 2025a): is a reasoning-based VQA framework that integrates LMMs with reinforcement learning to jointly perform video quality understanding and scoring, mimicking human percep-

tual judgment. It introduces a GRPO algorithm that incorporates three VQA-specific reward mechanisms, including bell-shaped regression, pairwise ranking, and temporal consistency, to enhance reasoning quality and prediction accuracy. Experiments show that VQAThinker achieves state-of-the-art performance, strong out-of-distribution generalization, and superior explainability in both quality scoring and distortion interpretation tasks.

12. **Q-Align** (Wu et al., 2023c) : a MLLM-based method that trains LMMs for image and video quality assessment by emulating human raters' use of discrete text-defined rating levels, enabling unified evaluation across IQA, IAA, and VQA tasks while achieving state-of-the-art performance.

13. **MinimalisticVQA** (Sun et al., 2024): a ultra-simple blind video quality assessment model designed to analyze and benchmark VQA datasets. They are built using only basic components: a video preprocessor (for spatiotemporal downsampling), a spatial quality analyzer, an optional temporal quality analyzer, and a quality regressor, all implemented with the simplest possible methods. Despite their simplicity, MinimalisticVQA models reveal fundamental issues in existing VQA datasets, such as the easy dataset problem, and provide insights for designing more robust and generalizable BVQA models.

14. **DeQA-Score** (You et al., 2025): a MLLM-based method that leverages MLLMs for precise image quality score regression by discretizing continuous score distributions into soft labels, preserving inter-image relationships, and employing a Thurstone-based fidelity loss to co-train across multiple IQA datasets, achieving stable performance aligned with human annotations.

15. **Q-Insight** (Li et al., 2025): an MLLM-based approach that leverages reinforcement learning via GRPO to jointly optimize score regression and degradation perception, enabling visual reasoning for image quality understanding under limited supervision. However, its reward design primarily focuses on absolute score regression, which often leads to overfitting specific distributions and poor robustness when generalized to unseen datasets.

16. **VisualQuality-R1** (Wu et al., 2025a): an MLLM-based reasoning-driven no-reference image quality assessment model that employs reinforcement learning to rank for human-aligned, context-aware evaluation. Its pairwise ranking formulation emphasizes relative ordering without sufficient score calibration, resulting in score distributions that deviate from human perceptual preferences.

17. **Qwen2.5-VL-7B** (Bai et al., 2025): is a pre-trained MLLM baseline used for comparison, providing strong multi-modal representation capabilities without task-specific fine-tuning.

18. **UnifiedReward-Think** (Wang et al., 2025a): a unified multimodal reward model that extends UnifiedReward by incorporating explicit long CoT reasoning across both visual understanding and generation tasks. It employs a three-stage optimization process, including CoT reward initialization, rejection sampling-based generalization fine-tuning, and GRPO reinforcement refinement. By leveraging both explicit and implicit CoT reasoning, it achieves robust alignment with human perceptual judgments across diverse tasks.

19. **Our PreResQ-R1**: a Preference-Response Disentangled Reinforcement Learning framework for fine-grained and interpretable image quality assessment. It unifies absolute score regression and relative ranking consistency within a single reasoning-driven optimization paradigm. Specifically, PreResQ-R1 decomposes the reward into two complementary branches: a response-based component that enhances intra-sample coherence across multiple reasoning trajectories, and a preference-based component that preserves inter-sample perceptual ordering through magnitude-aware pairwise and triplet comparisons. Trained with GRPO and a two-stage Exploration-to-Stability fine-tuning strategy, PreResQ-R1 enables MLLMs to perform human-aligned, chain-of-thought reasoning about visual fidelity, achieving both high correlation accuracy and robust cross-domain generalization.