

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: <https://orca.cardiff.ac.uk/id/eprint/190279/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Zhang, Zicheng, Li, Wenzhe, Chen, Liang, Wang, Junying, Wen, Farong, Guo, Yiji, Shen, Ye, Yan, Qihang, Li, Chunyi, Zhang, Wenlong, Zhou, Wei, Liu, Xiaohong, Min, Xiongkuo and Zhai, Guangtao 2026. Icme 2026 Grand Challenge on Scientific Image Quality Assessment. Presented at: International Conference on Multimedia and Expo Workshops, Bangkok, Thailand, 5-9 July 2026. Proceedings of ICMEW 2016. IEEE, 10.1109/icmew69720.2026.11720703

Publishers page: <https://doi.org/10.1109/icmew69720.2026.11720703>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See <http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.



ICME 2026 Grand Challenge on Scientific Image Quality Assessment

Zicheng Zhang¹, Wenzhe Li¹, Liang Chen¹, Junying Wang¹,
Farong Wen¹, Yiji Guo¹, Ye Shen¹, Qihang Yan¹, Chunyi Li¹, Wenlong Zhang¹,
Wei Zhou², Xiaohong Liu³, Xiongkuo Min³, Guangtao Zhai^{1,3}
Shanghai Artificial Intelligence Laboratory¹, Cardiff University², Shanghai Jiao Tong University³
Project Page: <https://siqa-competition.github.io/>

Abstract—The rapid integration of artificial intelligence into scientific workflows has increased the need for reliable Scientific Image Quality Assessment (SIQA). This paper presents the ICME 2026 Grand Challenge on SIQA, which benchmarks the capabilities of current large models to evaluate scientific imagery across four key dimensions: Validity, Completeness, Clarity, and Conformity. The challenge comprises two tracks: SIQA-U (Understanding), emphasizing multi-dimensional reasoning, and SIQA-S (Scoring), focusing on alignment with expert human judgment. Using the SIQA dataset, which spans diverse scientific domains and contains high-quality expert annotations, the competition attracts participation from seven international teams. Our analysis of submitted methods reveals a notable trend toward supervised fine-tuning and ensemble-based MLLM architectures. While top-performing models exhibit strong correlation with human perceptual scores, the results underscore a persistent gap in deep scientific reasoning.

Index Terms—Quality Assessment, Scientific Image, Understanding & Scoring

I. INTRODUCTION

Image Quality Assessment (IQA) is a fundamental task in computer vision, aiming to evaluate the perceptual quality of images in a manner consistent with human judgment [1]–[3]. Over the past decade, numerous IQA challenges have been organized, significantly advancing the field. These include the NTIRE series for perceptual image enhancement [4]–[6] and various benchmarks focusing on blind IQA for natural scenes as well as AI-generated content [7]. While these efforts have matured for common, assessing scientific imagery (such as microscopy, medical scans, and technical plots) remains an under-explored frontier.

Unlike common images, where quality is often defined by signal-level distortions like noise or compression [8], Scientific Image Quality Assessment (SIQA) requires a dual-track evaluation of both perceptual clarity and scientific integrity. Traditional metrics like Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [9], and even deep perceptual metrics like LPIPS [10], often fail to correlate with expert opinions in scientific domains because they do not account for factual accuracy or technical completeness. Consequently, models that perform well on standard benchmarks may produce clean scientific images that are misleading or lack critical metadata like scale bars and proper labeling.

Recent advancements in Multimodal Large Language Models (MLLMs) have introduced new possibilities for knowledge-driven quality assessment. However, existing general-purpose MLLMs often struggle to bridge the gap between general visual reasoning and specialized scientific mastery. To address these limitations, we organize the **ICME 2026 Grand Challenge on Scientific Image Quality Assessment**. This challenge provides a benchmark for evaluating AI’s ability to interpret images across four dimensions: validity, completeness, clarity, and conformity. By moving beyond simple signal-based metrics, this competition aims to catalyze the development of highly trustworthy evaluators for the global scientific community.

II. CHALLENGE OVERVIEW

A. Task Definition

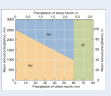
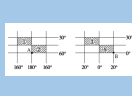
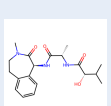
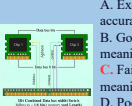
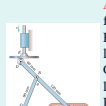
The ICME 2026 Grand Challenge on SIQA is formulated as a dual-track competition designed to evaluate MLLMs across the spectrum of *Perception-to-Knowledge* assessment. Unlike standard IQA which focuses on a single Mean Opinion Score (MOS), SIQA requires models to decouple their reasoning:

- 1) **Track 1: SIQA-U (Understanding)**: This track uses multiple-choice questions (MCQs) to assess a model’s deep understanding of scientific images. Grounded in objective, fact-based judgments, SIQA-U ensures reproducible and reliable evaluation.
- 2) **Track 2: SIQA-S (Scoring)**: This track focuses on the regression capability of models. Participants must predict continuous quality scores that align with the expert-consensus scores. The objective is to push models beyond mere ranking toward precise, human-aligned quantification of scientific validity.

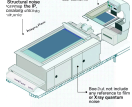
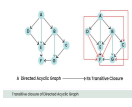
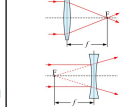
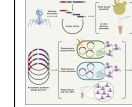
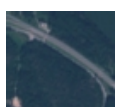
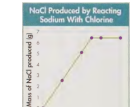
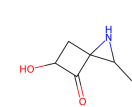
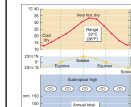
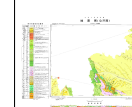
B. Dataset

The competition utilizes the SIQA dataset [11], specifically curated to ensure a highly-trustworthy ground truth (*overview shown in Fig. 1*). The details of the dataset are as follows:

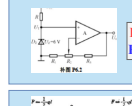
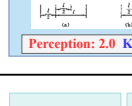
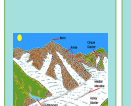
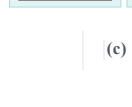
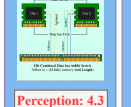
- **Domain Diversity**: The dataset spans multiple disciplines, including microscopy (SEM/TEM), medical imaging, and data-driven plots (e.g., heatmaps, scatter plots, and multi-panel figures).

SIQA (U)	Scientific Completeness (Comp.)	Scientific Validity (Val.)	Cognitive Clarity (Clar.)	Disciplinary Conformity (Conf.)
Yes/No	Question: Does the image <u>include a legend</u> for the anatomical illustration?  A. Yes B. No	Question: Does the depicted acetaldehyde <u>correctly</u> show a carbon oxygen double bond?  A. Yes B. No	Question: Does the diagram's labeling follow best practices to ensure <u>interpretive clarity</u> in ecological visualizations?  A. Yes B. No	Question: Does the angle labeling in the diagram <u>follow</u> international mathematical <u>standards</u> for communication?  A. Yes B. No
What	Question: Which key element is <u>missing</u> from the climatic classification diagram that affects completeness?  A. Tick marks on the axes B. Color coding for regions C. Axis labels for precipitation unit D. A title explaining the climate classification system	Question: Which aspect of the right diagram <u>contradicts</u> standard geographic coordinate representation conventions?  A. Tick marks on the axes B. Color coding for regions C. Axis labels for longitude and latitude units D. A title explaining the climate classification system	Question: Which issue negatively affects the image's ease of understanding?  A. Symbols like a microscope B. Consistent grayscale tones with selective color highlights C. Spelling errors in the title D. Clear sequential layout with labeled sections	Question: Which aspect of the image <u>aligns with</u> chemistry's <u>disciplinary</u> conventions for illustrating solutions?  A. Using green spheres to represent solute particles B. Omitting units for particle count C. Lacking temperature/pressure details D. Not specifying the solvent type
How	Question: How <u>complete</u> and <u>concise</u> is the scientific information presented in the image?  A. Excellent: All key info, no clutter; instantly clear. B. Good: Mostly complete; minor omissions or redundancy, still clear. C. Fair: Missing key details or has clutter; harder to interpret. D. Poor: Critical info missing and/or very cluttered; interpretation unreliable.	Question: To what extent <u>do typographical errors</u> in the image's text compromise its knowledge correctness?  A. Excellent: No typos; all content is accurate and unambiguous. B. Good: Minor typos; no impact on meaning or accuracy. C. Fair: Typos cause ambiguity; meaning is unclear but not wrong. D. Poor: Typos create factual errors or mislead; knowledge is seriously incorrect.	Question: How effectively does the image's layout support <u>understanding</u> of the mechanical linkage system?  A. Excellent: Clear, well-organized, fully labeled; easy to understand. B. Good: Clear and mostly well-labeled; minor room for improvement. C. Fair: Has key info but unclear labels; requires extra effort. D. Poor: Disorganized, hard to read, missing labels, and cluttered; very hard to understand.	Question: How well does the image to <u>disciplinary conventions</u> for instructional materials in terms of visual quality?  A. Excellent: All key map elements present and correct; clear standard. B. Good: Mostly complete; only a small element missing. C. Fair: Several key elements missing or unclear; harder to read. D. Poor: Critical flaws in legend, or geography; misleading or confusing.

(a) the example of SIQA-U about 4 dimension and 3 question types

SIQA (S)	<i>Bad</i>	<i>Poor</i>	<i>Fair</i>	<i>Good</i>	<i>Excellent</i>
Knowledge Driven	Knowledge: 1.1 Perception: 2.0 	Knowledge: 2.0 Perception: 3.1 	Knowledge: 3.2 Perception: 3.7 	Knowledge: 4.1 Perception: 3.2 	Knowledge: 5.0 Perception: 4.9 
Perception Driven	Perception: 1.3 Knowledge: 1.0 	Perception: 2.1 Knowledge: 3.1 	Perception: 3.2 Knowledge: 4.0 	Perception: 4.0 Knowledge: 4.1 	Perception: 4.9 Knowledge: 4.9 

(b) examples of different levels about SIQA-S

Knowledge Driven  Perception: 2.7 Knowledge: 4.0  Perception: 2.0 Knowledge: 3.7	 Perception: 3.8 Knowledge: 4.2  Perception: 4.9 Knowledge: 4.9
 Perception: 1.3 Knowledge: 1.0  Perception: 1.6 Knowledge: 1.0	 Perception: 4.3 Knowledge: 2.2  Perception: 3.9 Knowledge: 4.2
Perception Driven	

(c) Perception vs. Knowledge in SIQA-S.

Fig. 1. The overview of the SIQA challenge dataset.

- **Annotation Protocol:** Every image underwent a three-stage expert-in-the-loop review process. Initially, images were scored by domain experts; subsequently, a meta-review resolved discrepancies to ensure inter-rater reliability. The final dataset consists of a training set, a development set for local validation, and a strictly sequestered test set used for the final leaderboard.
- **Quality Dimensions:**
- **SIQA-S (Scoring):** Each entry is evaluated along two core quality dimensions: *Perception* and *Knowledge*, providing a structured assessment of model performance in scientific image quality evaluation.

C. Evaluation Protocol

To reflect the multidisciplinary nature of Scientific Image Quality Assessment, we employ a combination of semantic reasoning benchmarks and statistical correlation metrics.

1) *Track 1: SIQA-U (Understanding):* Performance in Track 1 is measured via structured visual question answering (VQA) across three distinct cognitive tiers. The evaluation prioritizes deep reasoning over binary verification by applying the following weighting to the accuracy scores:

- **Yes/No ($ACC_{yes/no}$):** Binary verification of factual or structural conditions.
- **What (ACC_{what}):** Multiple-choice comprehension of scientific entities and relationships.
- **How (ACC_{how}):** Complex quality judgments regarding completeness and disciplinary conventions.

The final score for Track 1 is calculated as:

$$S_{track1} = 0.2 \cdot ACC_{yes/no} + 0.3 \cdot ACC_{what} + 0.5 \cdot ACC_{how} \quad (1)$$

2) *Track 2: SIQA-S (Scoring):* Track 2 evaluates the model's ability to predict continuous quality scores along two complementary axes: **Knowledge-Driven** (Factual cor-

rectness) and **Perception-Driven** (Visual quality). Models are required to predict these scores directly from the visual input without auxiliary text prompts.

For each dimension $d \in \{\text{Know.}, \text{Percep.}\}$, we calculate a dimension-specific score using the Spearman Rank Correlation Coefficient (SRCC) and Pearson Linear Correlation Coefficient (PLCC):

$$\text{Score}(d) = \max \left(\frac{\text{SRCC}(d) + \text{PLCC}(d)}{2}, 0 \right) \times 100 \quad (2)$$

The final ranking metric for Track 2 is the arithmetic mean of the two dimensional scores:

$$S_{\text{track2}} = \frac{\text{Score}_{\text{Know.}} + \text{Score}_{\text{Percep.}}}{2} \quad (3)$$

By separating these scores, the challenge identifies models that may excel at perceptual alignment while lacking the knowledge accuracy required for reliable scientific assessment.

D. Baseline Models

To establish performance boundaries for current image quality assessment (IQA) methods, we consider a diverse set of baseline models:

- **Traditional IQA:** We include NIQE [8] as a representative hand-crafted, no-reference metric. This baseline reflects the ability to evaluate the naturalness and perceptual quality of scientific images without relying on learning-based models.
- **Large-Model Evaluators:** We evaluate the performance of recent large-model-based IQA approaches, including CLIP-IQA [12] and Q-Align [13] which leverage pre-trained vision-language models for quality assessment.

III. CHALLENGE RESULTS

A. Leaderboard

The ICME 2026 Grand Challenge on SIQA attracts participation from seven international teams. All submissions are evaluated on a sequestered test set to ensure robustness and generalizability. The challenge features two complementary tracks: Track 1 (SIQA-U) evaluates high-level reasoning across Yes/No, What, and How questions, while Track 2 (SIQA-S) focuses on continuous quality scoring along the Knowledge and Perception dimensions.

For Track 1, evaluation is based on the Weighted Accuracy (Acc_w) across the three question types (*results shown in Table I*). Analysis shows that most teams achieve strong alignment with perceptual criteria, but the “How” questions (*requiring nuanced quality judgments*) remain the primary factor differentiating top-performing models from the baselines. Table II presents the results for Track 2, highlighting the performance of the top teams relative to established baselines. The results show that leading teams substantially outperform traditional and zero-shot methods in both perceptual and knowledge-driven metrics, demonstrating the effectiveness of large-model-based approaches for continuous scoring.

TABLE I
SIQA-U LEADERBOARD (UNDERSTANDING).

Team	Yes/No	What	How	Final Score
DB_Team (Winner)	53.95	78.86	34.87	51.88
Tomcat	58.68	83.71	28.21	50.95
leitinggaba	53.95	81.43	31.28	50.86
DoubleY	55.79	75.43	27.18	47.38
YWJC	46.05	70.29	28.21	44.40

B. Team Solutions

1) *Team Tomcat:* For Track 1 (SIQA-U), the team reformulated the multiple-choice task as a candidate scoring and intra-question ranking problem. They employed the Qwen3.5-2B vision-language model as the backbone, with inputs comprising the image, question, and a task-aware prompt encoding image category, quality aspect, and question-type-specific instructions. A scalar score for each option was generated by extracting the last hidden state at the answer position and passing it through an MLP head. Parameter-efficient fine-tuning was performed using LoRA (rank $r = 8$, $\alpha = 32$) on attention and MLP layers, while the visual encoder remained frozen. Training combined a listwise multi-classification loss (intra-question softmax cross-entropy) and a pairwise margin ranking loss, with emphasis on “How” questions. For Track 2 (SIQA-S), they decomposed the dual-dimensional scoring into two conditional sub-tasks: Knowledge (scientific validity and completeness) and Perception (cognitive clarity and disciplinary conformity). Using the Qwen3.5-9B model, each sample was duplicated into two training instances—one per dimension—with dedicated prompts to guide focus. Ground-truth scores were mapped to five soft ordinal labels (Bad to Excellent), and the model predicted a probability distribution over these classes via an MLP head; the final continuous score was computed as the expectation. Training used a weighted combination of soft-label cross-entropy and mean squared error loss ($\lambda = 0.2$), with LoRA-based fine-tuning and a frozen visual encoder.

2) *Team leitinggaba:* For Track 1 (SIQA-U), the team fine-tuned the Qwen3-VL-8B-Instruct model via Supervised Fine-Tuning (SFT). To improve robustness, they adopted a Multi-Checkpoint Majority Voting strategy during inference, aggregating predictions from multiple training checkpoints to determine the final answer. For Track 2 (SIQA-S), they implemented a Decoupled LoRA Strategy based on InternVL3.5-8B within the LLaMA-Factory framework. The dataset was split into knowledge and perception subsets, and two independent LoRA adapters (rank $r = 16$) were trained—one for scientific validity/completeness and another for visual/disciplinary quality. During inference, Prompt Ensembling was used: each image was evaluated with four prompt variants, and scores were averaged across prompts per dimension.

3) *Team DoubleY:* The team participated in both tracks with SciQNet, a two-stage multimodal adaptation framework built on Qwen3-VL-2B. For Track 2 (SIQA-S), they first

TABLE II
SIQA-S LEADERBOARD (SCORING).

Model/Team	Perception		Knowledge		Final Score
	SRCC	PLCC	SRCC	PLCC	
<i>Baselines:</i>					
NIQE (Zero-shot)	0.3452	0.2353	0.4477	0.4101	35.92
Q-Align (Zero-shot)	0.7490	0.7629	0.2852	0.4001	54.90
InternVL3.5 (FT)	0.8573	0.8812	0.9151	0.9376	89.75
<i>Participants:</i>					
Team Tomcat (Winner)	0.9135	0.9253	0.9282	0.9527	92.99
Team DoubleY	0.9069	0.9271	0.9130	0.9414	92.21
Team leitinggaba	0.9031	0.9205	0.9038	0.9379	91.63
Team YWJC	0.8915	0.9133	0.8824	0.9147	90.05
Team NJU_SIQA_Team	0.8763	0.9071	0.8514	0.8948	88.24
Team DB_Team	0.8352	0.8647	0.9097	0.9065	87.90
Team CSQA	0.8462	0.8734	0.8571	0.8860	86.57

conducted domain-adaptive pretraining on a filtered subset of scientific document images from the M-Paper dataset, followed by task-specific fine-tuning on SIQA-S. The task was cast as instruction tuning, where the model predicted discrete quality ratings (Bad–Excellent) for Perception and Knowledge separately. A continuous score was derived from the logits over rating tokens and supervised via Huber loss alongside causal language modeling loss, enabling joint regression and generation learning. For Track 1 (SIQA-U), they framed the task as multiple-choice VQA and jointly trained on SIQA-U and SIQA-S data using a task-aware sampler, with the model outputting a single letter (A–D) as the answer.

4) *Team NJU_SIQA_Team*: For Track 2 (SIQA-S), the team proposed SIQAModel, a lightweight multi-feature fusion architecture. It uses a ConvNeXt-Base backbone to extract visual features and integrates textual content via EasyOCR. The fused multimodal representation is processed by two separate prediction heads to independently estimate Knowledge and Perception scores. The model was trained with a composite loss combining MSE, Pearson correlation, ranking, and a custom consistency term.

5) *Team DB_Team*: The team employed a Retrieval-Augmented Generation (RAG) framework leveraging a multimodal index to support reasoning with relevant reference cases. For Track 1 (SIQA-U), they constructed a hybrid index combining textual semantics (via text-embedding-3-large) and fine-grained visual patches (via ColPali). Retrieval involved hard filtering by scientific class, followed by multi-route semantic search—textual, text-driven visual, and (for “How” questions) image-to-image retrieval—with results fused via weighted Reciprocal Rank Fusion and re-ranked by category/type. Top retrieved examples were passed to a large language model with a step-by-step reasoning prompt. For Track 2 (SIQA-S), they built a visual-only index using ColPali embeddings. For each test image, top visually similar training samples were retrieved and used as quality anchors in the prompt to guide score prediction for both dimensions.

6) *Team YWJC*: For Track 1 (SIQA-U), the team adopted a training-free approach using a proprietary Multimodal Large Language Model (GPT-5.4) guided by Precision-Matched Dynamic One-Shot Prompting. A tiered retrieval system selected the most relevant single example based on class, type, and category match, which was then used in a one-shot prompt for answer generation. For Track 2 (SIQA-S), they utilized the vision encoder of InternVL3.5-38B as a frozen feature extractor and trained a lightweight two-layer regression head to directly predict continuous scores for factual (Knowledge) and perceptual (Perception) quality from image features.

7) *Team CSQA*: For Track 2 (SIQA-S), the team developed a no-reference image quality assessment model based on Swin-Large-Patch4-Window12-384. Their architecture features a dual-branch decoupled regression design that independently predicts Perception and Knowledge scores. Input images are resized to 384×384 and normalized. A 1536-d global visual feature from the frozen backbone is fused with a one-hot vector of distortion type and a vector of scientific priors. Two identical regression branches—each with LayerNorm, ReLU, Dropout, and Sigmoid—output normalized scores. An auxiliary distortion classification head is jointly trained to enhance feature discriminability via multi-task learning.

IV. CONCLUSION

The ICME 2026 Grand Challenge on Scientific Image Quality Assessment establishes a benchmark for evaluating AI models on both perception & knowledge-driven dimensions of scientific imagery. The results demonstrate that large-model-based approaches achieve strong performance in continuous scoring (Track 2) and exhibit notable reasoning capabilities in categorical understanding (Track 1). Despite these advances, the leaderboard highlights persistent challenges in capturing domain-specific completeness and nuanced quality judgments. These findings suggest that while current models effectively assess perceptual quality and general scientific structure, higher-level reasoning and factual correctness remain areas for future research.

REFERENCES

- [1] Zicheng Zhang, Junying Wang, Farong Wen, Yijin Guo, et al., “Large multimodal models evaluation: A survey,” *SCIENCE CHINA Information Sciences*, vol. 68, no. 12, pp. 221301–221369, 2025.
- [2] Zicheng Zhang, Junying Wang, Yijin Guo, et al., “Aibench: Towards trustworthy evaluation under the 45° law,” *Displays*, p. 103255, 2025.
- [3] Zicheng Zhang, Ziheng Jia, Chunyi Li, Yingjie Zhou, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai, “Towards versatile multimedia quality assessment for visual communications,” *Science China Information Sciences*, vol. 69, no. 2, pp. 122306, 2026.
- [4] Xin Li, Kun Yuan, Bingchen Li, Fengbin Guan, Yizhen Shao, Zihao Yu, Xijun Wang, Yiting Lu, Wei Luo, Suhang Yao, et al., “Ntire 2025 challenge on short-form ugc video quality assessment and enhancement: Methods and results,” in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 1092–1103.
- [5] Xiaohong Liu, Xiongkuo Min, Guangtao Zhai, Chunyi Li, Tengchuan Kou, Wei Sun, Haoning Wu, Yixuan Gao, Yuqin Cao, Zicheng Zhang, et al., “Ntire 2024 quality assessment of ai-generated content challenge,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6337–6362.

- [6] Xiaohong Liu, Xiongkuo Min, Qiang Hu, Xiaoyun Zhang, Jie Guo, Guangtao Zhai, Shushi Wang, Yingjie Zhou, Lu Liu, Jingxin Li, et al., "Ntire 2025 xgc quality assessment challenge: Methods and results," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 1389–1402.
- [7] Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin, "Agiqa-3k: An open database for ai-generated image quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 8, pp. 6833–6846, 2023.
- [8] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.
- [9] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE TIP*, vol. 13, no. 4, pp. 600–612, 2004.
- [10] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018, pp. 586–595.
- [11] Wenzhe Li, Liang Chen, Junying Wang, Yijing Guo, Ye Shen, Farong Wen, Chunyi Li, Zicheng Zhang, and Guangtao Zhai, "Siqa: Toward reliable scientific image quality assessment," *arXiv preprint arXiv:2603.06700*, 2026.
- [12] Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy, "Exploring CLIP for assessing the look and feel of images," .
- [13] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al., "Q-align: Teaching lms for visual scoring via discrete text-defined levels," *arXiv preprint arXiv:2312.17090*, 2023.