TIQE - Text to AI Generated Image Quality Evaluator

Kocherla Sai Kiran
Department of Artificial Intelligence
IIT Hyderabad
ai24mtech02003@iith.ac.in

Eswar Venkata Sai
Department of Artificial Intelligence
IIT Hyderabad
ai24mtech11007@iith.ac.in

R. Suraj Kumar
Department of Artificial Intelligence
IIT Hyderabad
ai22btech11022@iith.ac.in

Kota Dhana Lakshmi
Department of Artificial Intelligence
IIT Hyderabad
ai22btech11012@iith.ac.in

Abstract

Recent improvements in generative models have greatly enhanced text-to-image synthesis, allowing for the generation of very realistic images from text descriptions. Still, it is difficult to assess the quality of output images because image quality is subjective and matching visual content to input text is complicated. Text-to-image generation involves complex cognitive processes, creating distinct challenges for evaluating image quality, which is often optimized for professional and user-generated content. Traditional image quality metrics primarily address natural distortions but fail to effectively capture the alignment between the generated image and its input prompt. This paper presents TIQE (Text-to-Ai Generated Image Quality Evaluator), a new approach to evaluate the quality of text-generated images. The proposed framework aims to leverage both perceptual quality metrics and semantic alignment techniques to provide a comprehensive evaluation of the generated images, assessing both their visual quality and their alignment with the given prompt.

1. Introduction

Text-to-image synthesis is currently a rapidly expanding field of artificial intelligence, driven by advances in deep learning models capable of generating high-quality images from text descriptions. This capability has tremendous potential in numerous applications, including digital art, content creation, and assistive technology. Despite these advances, evaluating the quality of synthesized images remains a challenging task. Traditional evaluation methods typically involve the use of subjective human judgments, being slow and not consistent. Metrics for automated evaluation, although objective, cannot be guaranteed to catch semantic coherence between the input text and the generated image.

Advancements in machine learning and computer vision have revolutionized multimedia communications, particularly with the rise of AI-Generated Content (AIGC). Generative models such as GANs, VAEs, LLMs, and diffusion models are now widely used to create realistic and artistic images, videos, and 3D scenes.However, visual quality assessment (IQA/VQA) for AI-generated and AI-enhanced content (AIGC/AIEC), emphasizing both the advancements and challenges in evaluating AI-generated media. Traditional quality assessment methods, such as SSIM[17] and VMAF[9], struggle to accurately capture the distortions present in AI-generated images and videos. To address this, deep learning-based methods like LPIPS[5], DISTS[1], and CLIP-IQA[14] have been introduced, offering improved alignment with human perception. However, conventional AI-specific metrics, such as FID and IS, often fail to fully capture the nuances of perceptual quality, leading to the development of more sophisticated benchmarks like PickScore[6], and VBench[3]. Subjective quality assessment, relying on human ratings such as Mean Opinion Scores (MOS), remains essential for evaluating AIGC. However, as AI-generated content evolves, existing methodologies face limitations in assessing aesthetics, naturalness, and the uncanny valley effect. Vision-language models (VLMs) offer promising solutions, but they require enhancements in abstraction levels to incorporate commonsense reasoning.

To overcome these limitations, we want to present TIQE (Text-to-AI Generated Image Quality Evaluator), a holistic system that can produce objective, consistent, and scalable quality estimates for text-to-image models.

2. Literature Survey

In this paper, they have focused on evaluating the visual quality of AGI, which can be used to filter high-quality images from generation systems. The traditional DNN-based models for IQA have been designed for natural images to target issues such as blur, noise, and other forms of degradation that might not work with AGIs. Although re-training existing IQA models on AGI datasets might increase the scores of IQA metrics, as the IQA model has trained on AGI datasets, it becomes biased to AI-generated images, which results in high scores and not to the naturalness or semantic correctness in the AGIs. Handcrafted models like BRISQUE, ILNIQE, and NIQE use Natural Scene Statistics (NSS) to detect distortions but struggle with real-world variations. Deep learning models, such as DBCNN and HyperIQA, leverage CNNs to handle both synthetic and authentic distortions, while Vision Transformer-based models like IQT, MUSIQ, TReS, and MANIQA improves quality assessment through multi-scale encoding and ranking-based learning. Despite these advancements, traditional IQA models still face challenges with AI-generated images prompt the use of Large Multi-Modality Models (LMMs) that integrate semantic understanding for more robust assessment.

Recent LMMs for IQA either operate independently or enhance performance when combined with DNNs. The MA-AGIQA[15] framework integrates MANIQA[18] as the DNN backbone, mPLUG-Owl2[19] for extracting semantic features via prompts, and a Mixture of Experts (MoE) to adaptively fuse quality and semantic cues for more focused image assessment.

Recent works like IPCE[10] leverage CLIP[12] encoders to align visual and textual embeddings, using text templates to represent similarity levels and refining predictions through regression for improved accuracy. Inspired by this approach, we adopt task-specific templates for coarse-grained analysis—alignment templates to evaluate semantic consistency and perception templates to assess visual quality. Additionally, we introduce fine-grained analysis by computing word-level similarities between the prompt and both global image features and local patch features. We also follow the same dataset splitting protocol as proposed in the IPCE framework.

3. Related Work

Traditional image quality assessment (IQA) methods, such as PSNR, SSIM, and BRISQUE, primarily target distortions like blur and noise but fall short when evaluating AI-generated images (AGIs), where semantic alignment with prompts is crucial. Deep learning-based models like WaDIQaM and HyperIQA improve robustness to distortions but remain text-agnostic.

To bridge this gap, vision-language models (VLMs) have been introduced for prompt-image alignment evaluation. CLIPScore and ImageReward, for example, use frozen CLIP encoders to score semantic relevance. However, these approaches often lack fine-grained quality differentiation and are sensitive to prompt phrasing.

Peng et al. [10] used the pre-trained **CLIP** [12] model to measure alignment by computing cosine similarity between handcrafted adverb-modified prompts and patch-level image features. The similarity scores are aggregated and mapped to quality predictions via regression. While efficient, this method suffers from fixed feature representations and limited adaptability to diverse prompt structures.

To address these shortcomings, Wang et al. [15] proposed a hybrid IQA framework that integrates MANIQA [18] for capturing low-level visual distortions and the frozen LMM mPLUG-Owl2 [19] for extracting high-level semantic alignment. The features from both models are fused through a gating network and passed to a regressor. Despite promising results, the frozen nature of mPLUG-Owl2 limits semantic adaptability.

Recent encoder selection studies have emphasized backbone architecture choices. For example, CLIP-based models using ViT backbones (e.g., ViT-B/32) tend to provide richer semantic representations than convolutional alternatives like RN50, due to the former's global receptive fields and self-attention mechanisms. Studies such as BLIP-2 [8] and ALIGN [4] demonstrate the benefits of combining frozen vision encoders (like ViT-G) with trainable language models for more effective cross-modal alignment. These models enable decoupled optimization and modular design, offering a better balance between semantic accuracy and efficiency.

Researchers are also focusing on similarity modeling approaches to overcome these limitations. DINOv2 [13] demonstrates that self-supervised vision transformers can learn robust semantic embeddings without paired text, offering an alternative for image-only encoders. Moreover, PromptAlign [16] introduces prompt refinement mechanisms that dynamically adjust textual queries to better match image content, highlighting the potential of adaptive prompt encoding. ImageBind [2] extends this idea by aligning vision, text, and audio modalities into a shared embedding space, further enriching alignment capabilities.

Overall, while VLMs offer powerful tools for text-image alignment, challenges remain around encoder selection, fine-tuning, and balancing generalization with task-specific accuracy. Exploring combinations of self-supervised vision encoders, adaptable prompt encoders, and multi-resolution similarity scoring could yield more robust and perceptually aligned IQA models in future work.

In this context, methods such as Prompt-AIGIQA [20] and PK-AIGIQA-4K [11] are particularly valuable. These

models are specifically designed to assess the quality of AI-generated images in a prompt-aware manner, emphasizing both semantic alignment and perceptual realism. Trained and evaluated on human-annotated datasets, they incorporate diverse prompt-image pairs to better reflect real-world evaluation conditions. Their architectures explore different encoder strategies—such as combining CLIP-based visual and textual embeddings or integrating adaptive prompt conditioning—to capture subtle differences in alignment and quality. Hence, encoder selection plays a vital role, as the capacity to represent both visual and semantic nuances directly impacts the model's ability to generalize across varied prompts and generative styles.

4. Methodology

4.1. Image-Processing and Embeddings

Initially, we acquired the archive of the AGIQA-3k[7] data set, which contains N images. And resized all images to (3,224,224).We next obtained CLIP image-encoder embeddings for each full image—yielding a matrix $F_I^{(r)} \in \mathbb{R}^{N\times 512}$ —for Coarse-Grained similarity and fine-grained similarity where we used coarse-grained similarity for alignment and preception and fine-grained for detailed quality understading which discussed more on respective sections. Also,each image was partitioned into 64 non-overlapping patches; these patches were independently encoded by the same CLIP model to produce $F_I^{(p)} \in \mathbb{R}^{(N\times 64)\times 512}$. we used $F_I^{(p)}$ for fine-grained similarity. Since all 64 patches originate from the same source image, we then averaged each group of 64 patch embeddings to reconstruct an image-level descriptor $\bar{F}_I^{(p)} \in \mathbb{R}^{N\times 512}$. we used for $\bar{F}_I^{(p)}$ coarse-grain similarity. Finally, we encapsulated this workflow in a single function that returns:

1. the resized-image embeddings $F_I^{(r)}$,
2. the raw patch embeddings $F_I^{(p)}$,
3. the averaged patch embeddings $\bar{F}_I^{(p)}$.

Embedding Set	Shape	Description
$F_I^{(r)}$	$(N, 512)$	Full-image embeddings
$F_I^{(p)}$	$(N \times 64, 512)$	All 64 patch embeddings
$\bar{F}_I^{(p)}$	$(N, 512)$	Per-image averaged patches

Table 1. Summary of image embeddings

4.2. Text Prompt Construction and Embeddings

To enable multi-granularity evaluation of AI-generated images, we construct two categories of text prompts for each image. The first category comprises **task-specific prompts** (T_{ts}), which are designed to assess both perceptual quality and alignment fidelity. For each initial prompt pt, we generate 12 textual variants using three linguistic strategies: *(i) adverb-based*, T_{ts}^{adv} = "A photo that {adv} matches {pt}", where $\{adv\} \in \{badly, poorly, fairly, well, perfectly\}$; *(ii) adjective-based*, T_{ts}^{adj} = "A photo of {adj} quality", where $\{adj\} \in \{bad, average, good\}$; and *(iii) antonym-based*, T_{ts}^{ant} = "{ant} photo", where $\{ant\}$ is an antonymic modifier (*bad*, *good*, *perfect*).

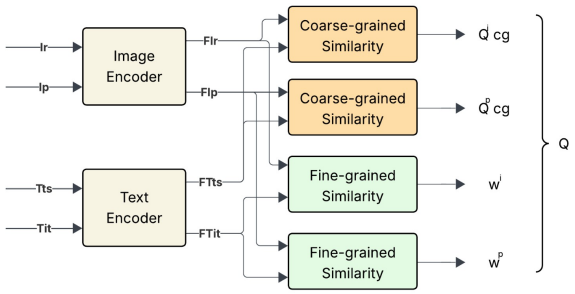

Figure 1. Model Architecture

Each T_{ts} is encoded using the CLIP text encoder to obtain a sentence-level embedding for coarse-grained similarity measurement.

The second category consists of the **initial prompts** (pt), which are the textual descriptions originally used to generate the images. Each prompt is tokenized into individual words indexed as FW_k, where $k \in \{1, 2, \ldots, K\}$. Word embeddings FW_k are extracted using the CLIP ViT-B/32 encoder, and all sequences are padded or truncated to a fixed length of 16, yielding a tensor of shape $[N, 16, 512]$ for N prompts. These are used for fine-grained token-level similarity evaluation.

This two-tiered embedding strategy—sentence-level via T_{ts} and token-level via FW_k—enables a comprehensive evaluation of both perceptual and semantic alignment in AIGIs.

4.3. Coarse-Grained Quality

The coarse-grained similarity measurement provides a fundamental framework for assessing AI-generated image quality by establishing semantic alignment between visual content and textual quality descriptions. This approach addresses two critical challenges in quality assessment: (1) the need for holistic quality evaluation beyond pixel-level distortions, and (2) the requirement for interpretable quality metrics that correlate with human perception.

Image-level similarity measures global alignment between the resized image and quality prompts, capturing overall composition and style adherence:

$$S_j^I = \frac{F_{I_r} \cdot F_{T_j}}{\|F_{I_r}\| \|F_{T_j}\|}, \quad j \in \{1, \ldots, L\} \tag{1}$$

Patch-level similarity assesses local quality through av-

eraged patch features, enabling detection of fine-grained artifacts that might be overlooked in global analysis:

$$S_j^P = \frac{\bar{F}P \cdot FT_j}{\|\bar{F}P\|\|FT_j\|} \tag{2}$$

Prompt Variant Selection. The textual quality descriptors F_{T_j} can be instantiated using any of three distinct prompt formulations:
where the template variables are populated from:

- **Adverb set** (adv): {`badly`, `poorly`, `fairly`, `well`, `perfectly`}
- **Adjective set** (adj): {`bad`, `poor`, `fair`, `good`, `perfect`}
- **Antonym set** (ant): {`bad`, `good`, `perfect`}

The dual-level similarity computation stems from the observation that human evaluators naturally assess images through both global impressions and local scrutiny. While traditional methods often focus exclusively on low-level features, this approach mirrors the hierarchical nature of human visual perception, where both macroscopic composition and microscopic details contribute to quality judgments.

Quality probability conversion transforms raw similarity scores into probabilistic distributions using softmax computation, accounting for the inherent subjectivity in quality assessment:

$$p_j^I = \frac{\exp(S_j^I)}{\sum_{l=1}^{L}\exp(S_l^I)}, \quad p_j^P = \frac{\exp(S_j^P)}{\sum_{l=1}^{L}\exp(S_l^P)} \tag{3}$$

This probabilistic framework was motivated by psychovisual studies showing that human quality ratings often cluster around discrete levels with smooth transitions between them. The softmax operation effectively models this behavior while maintaining ordinal relationships between quality grades.

Quality score regression converts probabilistic outputs into continuous scores suitable for regression tasks, preserving the interpretability of discrete levels while enabling fine-grained quality comparisons:

$$Q^{cg} = \frac{L}{L-1}\left(\sum j = 1^L j \cdot p^{j-1}\right), \quad * \in \{I, P\} \tag{4}$$

Final score integration implements an adaptive weighting mechanism that automatically balances global and local quality aspects based on their relative importance for each image:

$$Q_{cg} = \alpha Q_{cg}^I + (1-\alpha)Q_{cg}^P \tag{5}$$

The dynamic weighting reflects the empirical finding that different image types require different evaluation emphases, for instance, architectural renderings may demand stricter local quality control than impressionistic artworks. This flexibility represents a significant advancement over fixed-weight combination approaches prevalent in traditional quality assessment methods.we can make this trainable or set to 0 which considers only patches contribution or set to 1 which considers only resized original image contribution for score calculation.
In this manner this is applied to all the 2982 images on our dataset to get (2982,) tensor representing coarse grained quality scores for all the entries in the AGIQA-3k Dataset

4.4. Fine-Grained Similarity Measurement

While coarse-grained similarity measures the overall alignment between an image and its associated sentence-level prompt, fine-grained similarity captures more detailed token-wise correspondence between the image content and the keywords within the initial prompt.

Let the initial prompt be tokenized into K individual words, each represented by its CLIP embedding as FW_k, where $k \in \{1, 2, \ldots, K\}$. Given an image I_r, its global CLIP feature representation is denoted as F_{I_r}. The fine-grained similarity between the image and each word in the prompt is computed using the cosine similarity between F_{I_r} and FW_k. The aggregated image-level similarity score w_I is then obtained by averaging these word-level similarities:

$$w_I = \frac{1}{K}\sum_{k=1}^{K}\frac{F_{I_r} \cdot FW_k}{\|F_{I_r}\| \cdot \|FW_k\|}. \tag{6}$$

Similarly, we compute the patch-level similarity between the patch feature F_{P_n} and each word embedding FW_k. Let F_{P_n} denote the feature embedding of the n-th patch P_n within an image. For each patch, we calculate the mean cosine similarity with all K words in the initial prompt and take the maximum response across all patches to obtain the patch-level fine-grained score w_P:

$$w_P = \max_{n\in\{1,\ldots,N\}}\left(\frac{1}{K}\sum_{k=1}^{K}\frac{F_{P_n} \cdot FW_k}{\|F_{P_n}\| \cdot \|FW_k\|}\right). \tag{7}$$

Together, w_I and w_P capture both global and local semantic consistency between the generated image and the word-level components of its initial prompt. These fine-grained similarity scores are then fused with the coarse-grained similarity measures to provide a holistic assessment of AI-generated image quality.

We convert the fine-grained similarity scores into a quality score Q_{fg} by averaging the image-level (w_I) and patch-level (w_P) similarities, scaled by a factor L:

$$Q_{fg} = \frac{w_I + w_P}{2} \times L. \tag{6}$$

The final predicted quality score Q for an AIGI is obtained by combining both coarse-grained and fine-grained

components. Specifically, we use a weighted sum of the coarse-grained scores from the image and patch branches, Q_{cg}^{I} and Q_{cg}^{P}, respectively, along with the fine-grained score Q_{fg}:

$$Q = \alpha \cdot Q_{cg}^{I} + (1 - \alpha) \cdot Q_{cg}^{P} + Q_{fg}, \tag{7}$$

where $\alpha \in [0, 1]$ is a hyperparameter used to balance the contribution of the global image and patch-level coarse-grained components.

To train the model, we use the Mean Absolute Error (MAE) loss between the predicted quality Q and the ground-truth subjective quality score Q':

$$\mathcal{L} = |Q - Q'|. \tag{8}$$

5. Experiements

5.1. Implemtation Details

All experiments are performed on a PC equipped with an **NVIDIA A40 GPU**, using **PyTorch 1.12.0** and **CUDA 12.4**. In previous work(MID-TERM REPORT) , we implemented PK-AIGIQA and prompt-AIGIQA and through experimentation and results we found VIT-Base-32 model have better results. So We load the **ViT-B/32** as the backbone of our method, where input images are resized to 224×224. The model is optimized using the **AdamW** optimizer with a learning rate of 5×10^{-6} and a weight decay of 5×10^{-4}. Training is conducted for 20 epochs, and a **cosine annealing learning rate scheduler** is applied to gradually reduce the learning rate every 5 epochs. For coarse grained similarity all 3 task prompts i.e, (T_{ts}^{adv}, T_{ts}^{adj}, T_{ts}^{ant}) were calculated and averaged. The batch size is set to 16.

To ensure **reproducibility** and fairness in comparison we dividing AGIQA-3k dataset into training and testing sets in a 4:1 ratio.

5.2. Hyperparameters

- **Backbone Model:** ViT-B/32 (Vision Transformer)
- **Input Image Size:** 224 × 224
- **Optimizer:** AdamW
- **Learning Rate:** 5×10^{-6}
- **Weight Decay:** 5×10^{-4}
- **Learning Rate Scheduler:** Cosine annealing (updated every 5 epochs)
- **Number of Epochs:** 20
- **Batch Size:** 16
- **Dataset Split:** 80% training / 20% testing (4:1 ratio)
- **Experiment Runs:** 10 times (average results reported)

6. Results and Discussion

The analysis of CLIP-based similarity scores for prompt-image alignment using both ViT-B/32 and RN50 models revealed meaningful insights into how different backbones encode semantic alignment.As shown in Figure 2, the ViT-B/32 model yields higher and more dispersed CLIP similarity scores than RN50, indicating that its transformer-

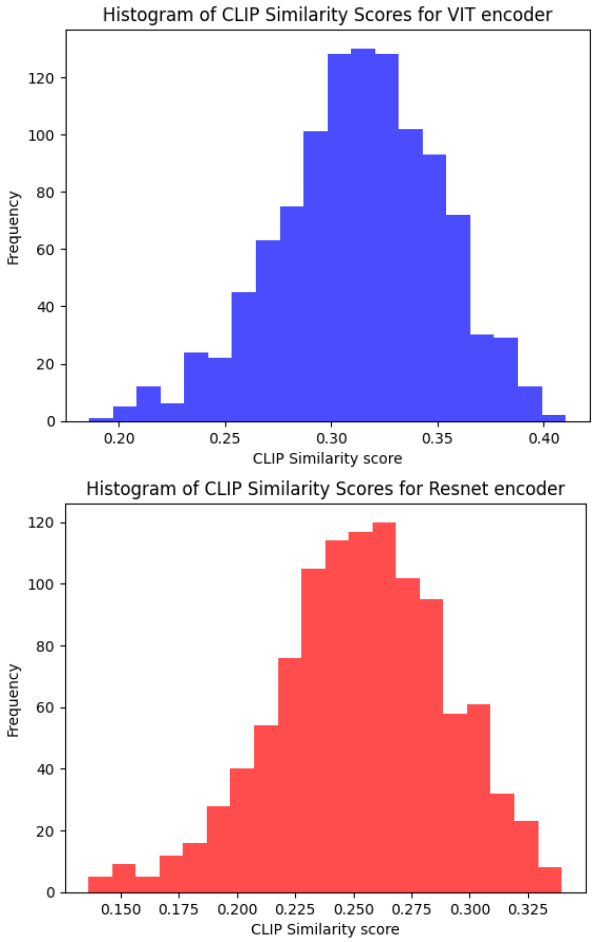

Figure 2. Distribution of prompt-image CLIP similarity scores. Left: ViT-B/32. Right: RN50.

Table 2. Comparison of SROCC and PLCC for Image Perception on AGIQA-3k dataset.

Method	SROCC	PLCC
PK-AIGIQA	0.8249	0.8773
Prompt-AIGIQA	0.8154	0.8807
Ours(TiQE)	0.8837	0.9117

Table 3. Comparison of SROCC and PLCC for Text-Image Alignment on AGIQA-3k dataset.

Method	SROCC	PLCC
PK-AIGIQA	0.6624	0.7801
Prompt-AIGIQA	0.608	0.7846
Ours(TiQE)	0.7746	0.8399

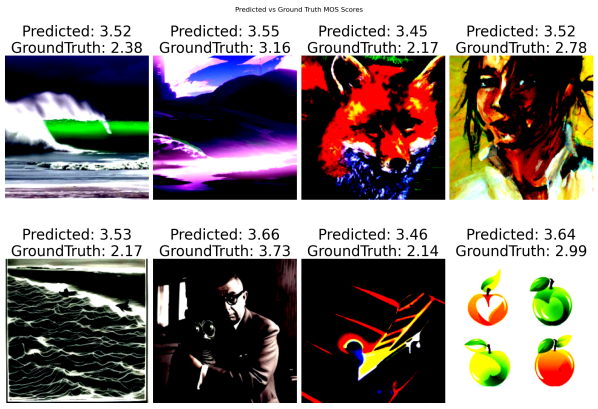

Figure 3. Illustration of some AIGIs with subjective score and prediction for Perception

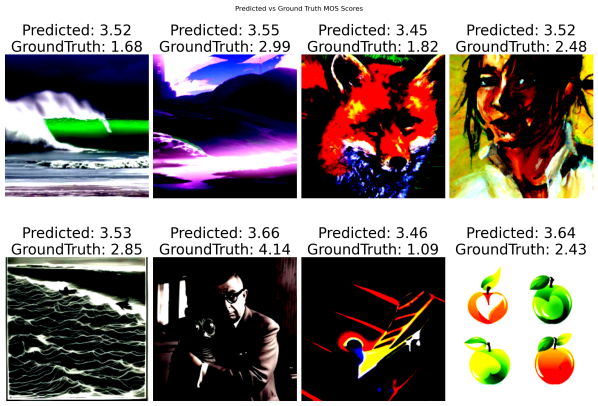

Figure 4. Illustration of some AIGIs with subjective score and prediction for Alignment

based architecture captures richer semantic associations. In contrast, RN50 produces a narrower distribution, reflecting more conservative estimations.

The results presented in Tables 2 and 3 for a comparative evaluation of the proposed method (Ours, denoted as TiQE) against two baselines—PK-AIGIQA and Prompt-AIGIQA—on the AGIQA-3k dataset. In terms of image perception (Table 2), TiQE outperforms both baselines, achieving the highest SRCC (0.8837) and PLCC (0.9117), indicating superior consistency and accuracy in quality prediction. Similarly, for text-image alignment assessment (Table 3), TiQE again shows marked improvements with an SRCC of 0.7746 and PLCC of 0.8399, surpassing the existing methods by a significant margin. These results demonstrate the effectiveness of TiQE in both perceptual quality estimation and multimodal alignment.

The perceptual quality evaluation results are illustrated in Figures 3 and 5, offering both qualitative and quantitative insights into the model's ability to predict human-judged Mean Opinion Scores (MOS). Figure 3 displays a set

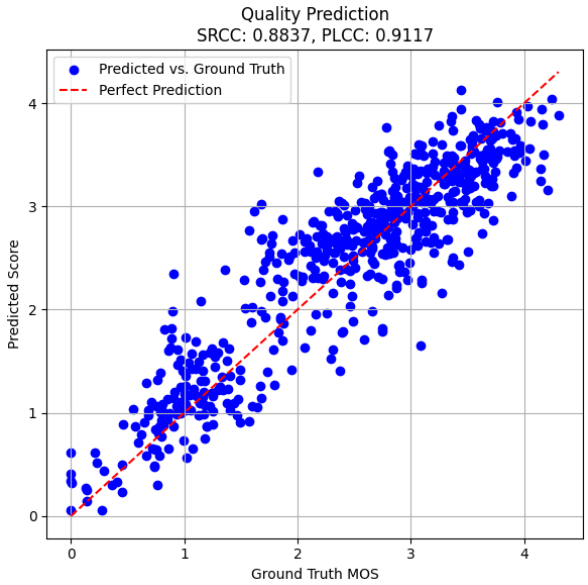

Figure 5. Predicted vs. ground truth MOS scores for Perception.

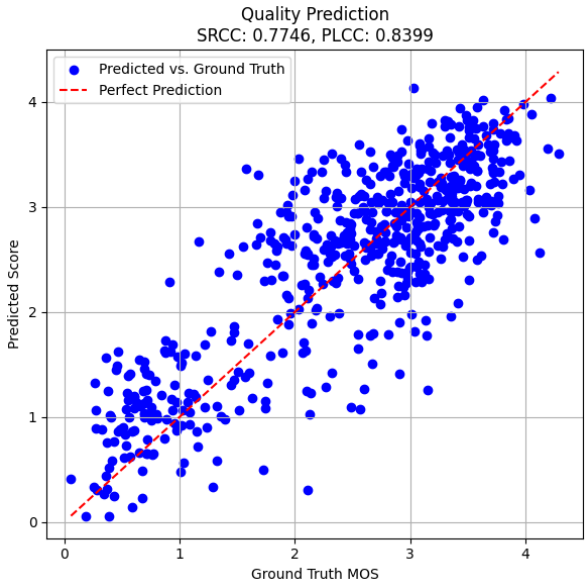

Figure 6. Predicted vs. ground truth MOS scores for Alignment.

of representative images, each annotated with the predicted and ground-truth MOS. These examples highlight how the model performs across a variety of content types and aesthetic styles. While the model demonstrates strong alignment in several cases—such as the image with a ground-truth score of 3.73 and a prediction of 3.66—there are also instances where perceptual quality is noticeably overestimated, particularly for images with lower MOS values (e.g., ground truth 2.17 vs. predicted 3.53). These discrepancies suggest that the model tends to favor visually appealing ele-

ments like color saturation or contrast, which may not fully reflect subtle degradations or artifacts perceived by human evaluators.

The alignment-based quality prediction results are presented in Figures 4 and 6, offering a comprehensive view of the model's effectiveness in capturing perceptual alignment between images and human quality ratings. Figure 4 showcases several test images with their predicted and ground-truth Mean Opinion Scores (MOS). The examples indicate that the model generally assigns moderately high predicted scores across the board, even when the actual ground-truth values are significantly lower. For instance, images with ground-truth scores as low as 1.09 and 1.68 are still predicted above 3.4, indicating a clear overestimation trend. This suggests that the alignment-based approach may be less sensitive to quality degradations that human raters readily detect, possibly because it emphasizes structural or semantic consistency more than fine-grained perceptual distortions.

7. Conclusion and Future Work

In this work, we proposed TiQE, a novel model for image quality estimation and text-image alignment, which outperformed existing methods like PK-AIGIQA and Prompt-AIGIQA on the AGIQA-3k dataset. TiQE achieved the highest SRCC and PLCC scores, demonstrating its effectiveness in predicting human-judged quality scores and capturing semantic relationships between images and prompts.Our analysis of CLIP-based similarity scores revealed that ViT-B/32 offers richer semantic alignments compared to RN50, which tends to be more conservative. This suggests that transformer-based architectures like ViT-B/32 may better capture multimodal relationships, explaining TiQE's superior performance.However, some observations highlighted areas for improvement.

The model overestimated the quality of images with lower MOS values, likely due to its emphasis on visually appealing features like color saturation. Additionally, the alignment-based predictions showed overestimation for images with lower ground-truth values, indicating a bias towards structural consistency over perceptual distortions.Future work could refine TiQE to better capture subtle perceptual distortions and improve its sensitivity to local quality degradations, potentially enhancing its performance on lower-quality images.

References

[1] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020. 1

[2] Rohit Girdhar, Alexander Kirillov, Joao Carreira, Saining Xie, and Ross Girshick. Imagebind: One embedding space to bind them all. *arXiv preprint arXiv:2305.05665*, 2023. 2

[3] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 1

[4] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zhifeng Parekh, Hieu Pham, Quoc V Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *International Conference on Machine Learning (ICML)*, 2021. 2

[5] Markus Kettunen, Erik Härkönen, and Jaakko Lehtinen. E-lpips: robust perceptual image similarity via random transformation ensembles. *arXiv preprint arXiv:1906.03973*, 2019. 1

[6] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 36652–36663, 2023. 1

[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*, pages 1–1, 2023. 3

[8] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C.H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2

[9] Marta Orduna, César Díaz, Lara Muñoz, Pablo Pérez, Ignacio Benito, and Narciso García. Video multimethod assessment fusion (vmaf) on 360vr contents. *IEEE Transactions on Consumer Electronics*, 66(1):22–31, 2020. 1

[10] Fei Peng, Huiyuan Fu, Anlong Ming, Chuanming Wang, Huadong Ma, Shuai He, Zifei Dou, and Shu Chen. Aigc image quality assessment via image-prompt correspondence. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 6432–6441, 2024. 2

[11] Bowen Qu, Haohui Li, and Wei Gao. Bringing textual prompt to ai-generated image quality assessment. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024. 2

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 2

[13] Hugo Touvron, Mathilde Caron, Andrei Bursuc, Matthieu Cord, Alaaeldin El-Nouby, Drew Hudson, Karel Lenc, Ishan Misra, Armand Joulin, Gabriel Synnaeve, and Herve Jegou. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[14] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2555–2563, 2023. 1

[15] Puyi Wang, Wei Sun, Zicheng Zhang, Jun Jia, Yanwei Jiang, Zhichao Zhang, Xiongkuo Min, and Guangtao Zhai. Large multi-modality model assisted ai-generated image quality assessment. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 7803–7812. ACM, 2024. 2

[16] Yujian Wang, Yuting Zhang, Kuang-Huei Lee, Shuicheng Yan, and Ziwei Liu. Promptalign: Prompt-based alignment for fine-grained text-to-image retrieval. *arXiv preprint arXiv:2301.12844*, 2023. 2

[17] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 1

[18] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment, 2022. 2

[19] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration, 2023. 2

[20] Jiquan Yuan, Fanyi Yang, Jihe Li, Xinyan Cao, Jinming Che, Jinlong Lin, and Xixin Cao. Pku-aigiqa-4k: A perceptual quality assessment database for both text-to-image and image-to-image ai-generated images. *arXiv preprint arXiv:2404.18409*, 2024. 2