

AAPMT: AGI Assessment Through Prompt and Metric Transformer

Benhao Huang, Shanghai Jiao Tong University

Abstract—The emergence of text-to-image models marks a significant milestone in the evolution of AI-generated images (AGIs), expanding their use in diverse domains like design, entertainment, and more. Despite these breakthroughs, the quality of AGIs often remains suboptimal, highlighting the need for effective evaluation methods. These methods are crucial for assessing the quality of images relative to their textual descriptions, and they must accurately mirror human perception. Substantial progress has been achieved in this domain, with innovative techniques such as BLIP and DBCNN contributing significantly. However, recent studies, including AGIQA-3K, reveal a notable discrepancy between current methods and state-of-the-art (SOTA) standards. This gap emphasizes the necessity for a more sophisticated and precise evaluation metric. In response, our objective is to develop a model that could give ratings for metrics, which focuses on parameters like perceptual quality, authenticity, and the correspondence between text and image, that more closely aligns with human perception. In our paper, we introduce a range of effective methods, including prompt designs and the Metric Transformer. The Metric Transformer is a novel structure inspired by the complex interrelationships among various AGI quality metrics.

Index Terms—Text-to-image models, AGIs, Quality evaluation, Text-Image correspondence(alignment), Image Authenticity, Human perception, Prompt design, Metric Transformer

I. Implementation of Model

A. Back Bone Model

Blip is an excellent model for Text-Image Matching task. We utilized a recent work Image Reward [1]. Based on this work, we try to fine-tune its pretrained model on AGIQA-3K dataset and AIGCIQA2023 dataset correspondingly to gain results.

B. Data Preparation

Regarding data preparation, we partitioned the entire dataset into two distinct subsets: 80% allocated for training and 20% for testing. Crucially, we adopted a strategy where images generated from the same prompt are grouped together. This approach adheres to the "Content Isolation" principle, ensuring that the results derived from the test set are more reliable. However, due to time constraints, we maintained a consistent random seed for the dataset split and did not conduct cross-validation experiments. We acknowledge this as a limitation and plan to incorporate cross-validation in future iterations of our

research to further validate our findings. But a simple robust test experiment will be included in Appendix.

C. Assess Text-Image Correspondence

Evaluating text-image correspondence can be conceptualized as a "Text-Image Match" problem: given an image and a paragraph of text as inputs, the model should output a score that quantifies the degree of match between the image and the text. This is precisely the model structure required for our task. Consequently, we employed the model provided by Image Reward¹, training it on our own dataset. In terms of hyperparameter selection, we largely adhered to the settings outlined in Image Reward[1]. Key hyperparameters are detailed in Table I. The reason for choosing for a higher number of epochs when training on the AGCIQA2023 dataset stems from the original model's performance on the AGIQA-3K dataset, where it was already competitive with numerous other Image Quality Assessment (IQA) models (SRCC = 0.874, PLCC = 0.834). However, a significant performance decline was observed when the model was tested on AIGCIQA2023, suggesting different properties between the two datasets. After conducting 10 epochs of training on AIGCIQA2023, we observed a significant improvement in the model's performance, but still suboptimal. Consequently, we increased the number of training epochs to 50, which we found sufficient to achieve satisfactory performance, as demonstrated in the figure 1.

D. Assess Image Quality

1) A Naive Trial: Utilizing a model adept in the text-image matching task offers a straightforward yet effective strategy: modify the input prompt to explicitly state "extremely high quality image, with vivid details" and train the model on this dataset. The underlying principle is simple: the higher the output score from the model, the greater the correspondence between the image and the concept of the prompt "extremely high quality image, with vivid details." This implies that the image in question is of high quality. Remarkably, this seemingly naive method yields surprisingly effective results. The performance of the model, as measured by Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank Correlation Coefficient (SRCC), is excellent. The results' values are detailed in the figure 2.

Benhao Huang is an undergraduate student from Shanghai Jiao Tong University. E-mail:hbh001098hbh@sjtu.edu.cn

¹<https://github.com/THUDM/ImageReward>

TABLE I
Hyperparameters Setting

| Model | dataset | Training Epochs | Batch Size | Learning Rate | Adam Beta 1 | Adam Beta 2 | Adam Eps |
|------------------|------------|-----------------|------------|---------------|-------------|-------------|----------|
| ImageReward-v1.0 | AGIQA-3K | 10 | 32 | 5e-06 | 0.9 | 0.999 | 1e-8 |
| ImageReward-v1.0 | AGCIQA2023 | 50 | 32 | 5e-06 | 0.9 | 0.999 | 1e-8 |

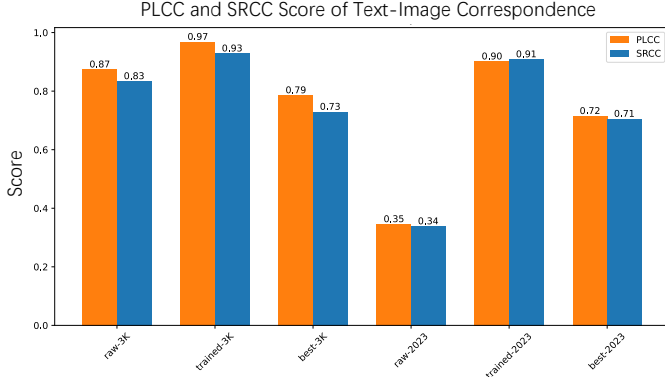


Fig. 1. Assess Text-Image Correspondence. '3K' signifies that the task is evaluated on the AGIQA-3K dataset, while '2023' indicates its evaluation on the AGCIQA2023 dataset. The term 'raw' refers to using the Image Reward model that has not been trained on the specific dataset, in contrast, 'trained' denotes the use of models that have undergone training on these datasets. Additionally, we also present the best scores as reported in references [2] and [3], corresponding to their respective datasets.

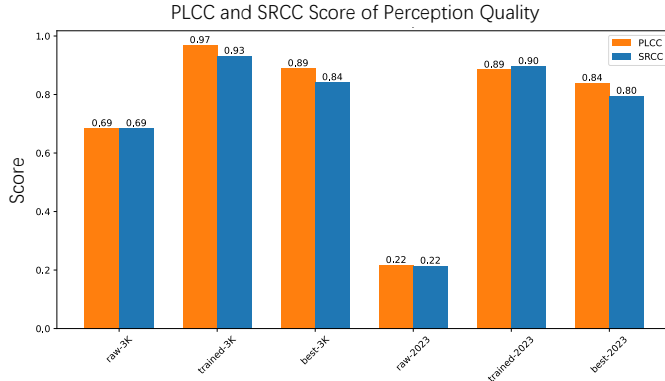


Fig. 2. Assess Image Quality by designing prompt. Same notations with figure 1.

2) Change in Prompt Content: Despite the initial success, we questioned whether this single prompt was optimal. To investigate further, we conducted additional experiments with three distinct prompts intended for image quality assessment. These prompts included: Prompt 1: "high quality image", Prompt 2: "extremely high quality image, with vivid details", and Prompt 3: "extremely high quality image, with high resolution". We reevaluated the model using these new prompts. The results, detailed in the table below, revealed some differences based on the specific prompt used. Notably, replacing "vivid details" with "high resolution" in Prompt 2 resulted in a decrease in both PLCC and SRCC scores. This finding suggests

that the AI model may place greater emphasis on "vivid details" than "high resolution" when assessing image quality.

TABLE II
PLCC and SRCC Score of Image Quality with Different Prompts

| | Prompt1 | Prompt2 | Prompt3 |
|------|---------|---------|---------|
| PLCC | 0.8812 | 0.8855 | 0.8594 |
| SRCC | 0.8941 | 0.8976 | 0.8780 |

Furthermore, the comparison between Prompt 1 and Prompt 2 shows a negligible difference, indicating that the phrase "high quality image" is the critical component in these prompts. Due to time constraints, we couldn't explore further possibilities, which would be an important avenue for future research.

E. Assess Image Authenticity

By following the same approach as used for assessing image quality, we selected the prompt "very authentic image" to evaluate the authenticity of images. Given that authenticity is not a metric included in the AGIQA-3K dataset, our evaluation was exclusively conducted on the AGCIQA2023 dataset. The results of this assessment are detailed in table III.

TABLE III
PLCC and SRCC Score of Image Authenticity

| | raw-2023 | trained-2023 | best-2023[3] |
|------|----------|--------------|--------------|
| PLCC | 0.1286 | 0.8985 | 0.6807 |
| SRCC | 0.1345 | 0.9082 | 0.6701 |

II. Assessing Multiple Metrics with a Single Model

So far, we have tried to use Image Reward model [1] to assess three metrics of AGI, which has a comparably excellent performance. Nevertheless, if we stick to this method, then we need to correspondingly train three models, namely models for quality, for authenticity as well as for text-image correspondence, which is quite inefficient and space-consuming. In this section, we want to come up with more novel ideas to further enhance the efficiency of AGI assessment. Since AGIQA-3K only has two metrics, we choose AGCIQA2023 dataset to carry out our further experiments. The following subsections elaborate on some different approaches we have attempted and the thought process we underwent in trying to solve this problem.

A. Second Training of the Model Initially Trained for a Different Task

A naive thought is that we could utilize the ability of a model which has been already trained for quality assessment to retrain it for another assessment task.

Algorithm 1 Second Training

- 1: Preparing the Dataset: Dividing the Dataset into Two Parts, with 20% for Testing and 80% for Training
 - 2: Model \leftarrow Initialize from Image Reward Model Parameters
 - 3: Train the Model for quality metrics
 - 4: Save the checkpoint $ckpt_{quality}$
 - 5: Reinitialize a Model and load the checkpoint $ckpt_{quality}$
 - 6: Train the Model for text-image correspondence metric
-

It's obvious that this method didn't save anytime, since the model in this method still needs to be trained multiple times for different metrics. However, we could gain some insightful points from its results. Here, we focus solely on retraining the model for text-image correspondence that was previously trained for quality assessment and the opposite case.

From table VI, we have following discoveries:

1. When retraining the trained model on a different task, its performance drops a lot on the original task. However, it's interesting to note that its performance, while reduced, remains better than that of a model which was never trained on the original task. For example, for AGCIQA2023-Align task, Align to Quality model still outperforms Trained for Quality.
2. Comparing to merely training on task A (quality or alignment), training on another task B (alignment or quality) then retraining the model on task A could further improve the quality.

From these phenomena, we could speculate that the optimal parameters space for quality task and text-image correspondence has relationship demonstrated by the following Venn diagram.

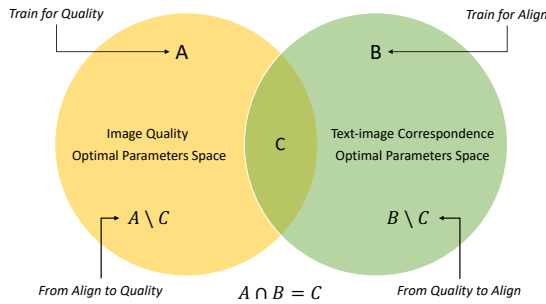


Fig. 3. Venn diagram of Optimal Parameters Space. The notations here share the meaning with table VI.

When retraining the model, which has been trained to assess quality, for alignment task, then its parameters space will shift from A to B , which will bring a drop

of performance on assessing image quality. However, since there is an overlap between two optimal parameters' space, a second training model on new task T could outperform the model trained only for task T , because in previous case, model already have a good knowledge encoded by the shared parameters space which is $A \cap B$. Moreover, we could speculate that when training only for one task, it's hard for model to embody the shared parameter space, which makes sense intuitively. Therefore, when using same training strategy and hyperparameters, a second training model, which goes through a second training for task A and first training for task B, could perform better on task A than those model merely trained for task A.

B. One Step Even Further: Metric Transformer

In subsection II-A, we explored various intriguing aspects of training models for different AGI assessment metrics. As previously noted, it appears that the optimal parameters for these metrics overlap to some extent. This observation leads to a plausible hypothesis: the metrics themselves might share semantic similarities. For example, an image with high text-image correspondence is likely to also be of higher quality, and vice versa. The one way we could utilize this phenomenon is to develop a model which could consider other metrics' influence when rating for a certain metrics. In order to make one metric assessment process have the information from other metrics, we develop Metric Transformer, which utilizes the advantage of self attention[4] mechanism.

Our model adopts the structure utilized in Image Reward[1] for text-image processing. We then replace the Multilayer Perceptron (MLP) layer of the original model with our proposed Metric Transformer. It's important to note that before feeding the text features into the Metric Transformer, we first pass them through a three-head transformer encoder. This initial step allows the model to learn the basic concepts of the three metrics from the training dataset.

These encoded features are then input into the Metric Transformer to calculate the scores for the three metrics correspondingly. If we denote these encoded features as EF , we then have:

$$\begin{aligned} Q_i &= W_Q^{i,T} \cdot EF \\ K_i &= W_K^{i,T} \cdot EF \\ V_i &= W_V^{i,T} \cdot EF \end{aligned} \quad (1)$$

Here W_Q^i, W_K^i, W_V^i is the parameters of models which is used to get Query, Key and Value matrices. Then we calculate the metric score as follows:

$$S_i = \sum_{j=1}^3 \text{SoftMax}(\frac{Q_i K_j^T}{\sqrt{d_k}} V_j) \quad (2)$$

Here d_k is the dimension of K_i . Then we trained this model with exactly the same setting of Image Reward model trained on AGCIQA2023, and derived the results shown in table V

TABLE IV

Results of Second Training. Align to Quality means training model for quality from a model which has been trained for assessing text-image correspondence task. Trained for Quality means merely training the model for assessing image quality task.

| Task | Align to Quality | | Quality to Align | | Trained for Quality | | Trained for Align | |
|--------------------|------------------|---------------|------------------|---------------|---------------------|--------|-------------------|--------|
| | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC |
| AGCIQA2023-Quality | 0.9173 | 0.9194 | 0.6199 | 0.6142 | 0.8855 | 0.8976 | 0.5316 | 0.5177 |
| AGCIQA2023-Align | 0.6641 | 0.6645 | 0.9145 | 0.9160 | 0.4652 | 0.5512 | 0.9018 | 0.9090 |

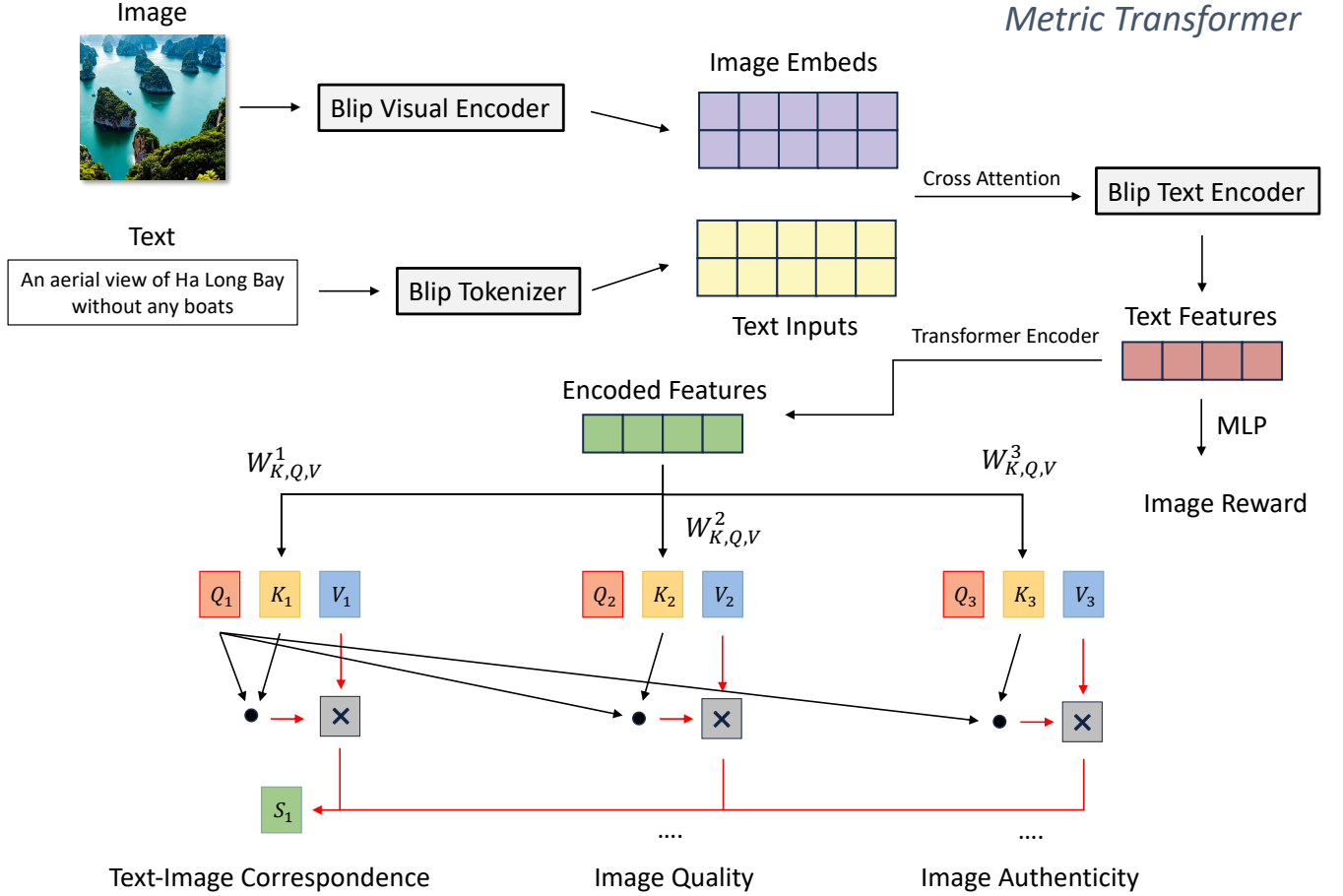


Fig. 4. Diagram of the Metric Transformer. It's worth noting that the primary distinction between our model and Image Reward lies in the design of the final layer. We employ a model similar to the transformer (which typically has only one set of $W_{K,Q,V}$). This design allows the model to evaluate multiple image metrics concurrently, delivering impressive performance as demonstrated in Table V.

TABLE V

Results of Assessing IQA Metrics Using Metric Transformer. Here the best score for a given metric on a task is designated by **RED**, while the second-best score is indicated by **BLUE**. The results demonstrate that the overall performance of the Metric Transformer is definitively on par with that of Image Reward. It's important to emphasize that the score for Image Reward is actually derived from three models specifically trained for different tasks, whereas the Metric Transformer involves only a single model. This point should be particularly considered when evaluating their performance.

| Task | Best in AIGCIQA2023 | | Trained Image Reward | | Metric Transformer | |
|-------------------------|---------------------|--------|----------------------|---------------|--------------------|---------------|
| | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC |
| AGCIQA2023-Quality | 0.8402 | 0.7961 | 0.8855 | 0.8976 | 0.8819 | 0.8852 |
| AGCIQA2023-Align | 0.7153 | 0.7058 | 0.9018 | 0.9090 | 0.9060 | 0.9068 |
| AGCIQA2023-Authenticity | 0.6807 | 0.6701 | 0.8985 | 0.9082 | 0.9112 | 0.9075 |

We also depict the loss trajectory during the training of the Metric Transformer in Figure 5. From the plot, it's evident that the loss continues to show a downward trend even after 50 epochs. However, considering the constraints of computational resources and time, we made the decision to halt training at this juncture.

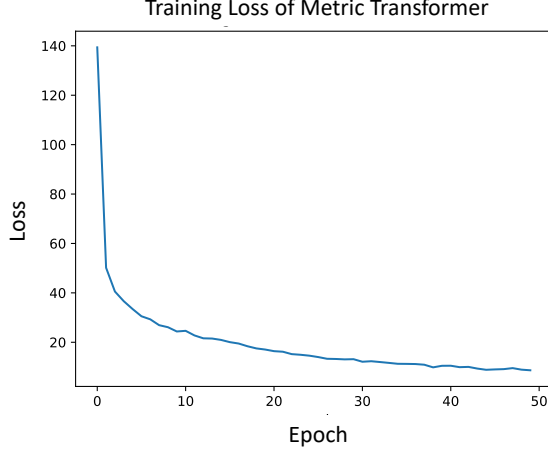


Fig. 5. Training Loss of Metric Transformer in 50 Epochs

III. Conclusion

In this paper, we leverage the model from previous work, Image Reward [1], to successfully evaluate multiple metrics relevant to AI-generated images. These metrics include perceptual quality, authenticity, and text-image correspondence. Beyond testing the naive method using pre-designed prompts, we further analyze the interplay between these metrics and hypothesize that they mutually influence each other. Drawing on these insights, we propose a novel model structure, the Metric Transformer, specifically designed to accelerate model training as well as improve model performance in AGI assessment. This model is capable of rating multiple metrics in a single run, displaying high correspondence with human evaluation scores. We delve deeper into our insights and render complementary experiments in the Appendix.

Appendix A

Test Model Performance with Different Seed

As we mentioned in I-B, we try to prove that the excellent performance of the model is not determined by some random factors. So we choose three seeds to initialize the dataset and model to ensure this point. We train and test on AGIQA-3K dataset.

Appendix B

More Discussion and Future Works

A. Design the Loss for Multiple Metrics

Regarding training a model for multiple metric assessment, we have thought that it's possible for us to design a dynamic loss, which is :

$$L = \sum_{i=1}^n \alpha_i \cdot Loss_i, \quad \text{where } \sum_{i=1}^n \alpha_i = 1 \quad (3)$$

Here α_i is called loss parameters, and $loss_i$ is the model's loss on metric i . During the training process, we update parameters $\alpha_i, i = 1, 2, 3, \dots, n$ as follows:

Algorithm 2 Updating Loss Parameters

```

1: Randomly Initialize  $\alpha_i$ , which satisfies  $\sum_i \alpha_i = 1$ 
2: previous  $\alpha_i \leftarrow \alpha_i$ 
3: for each step do
4:   for each metric  $i$  do
5:      $loss_i \leftarrow \text{model}(\text{input}, \text{metric}_i)$ 
6:   end for
7:    $\alpha'_i \leftarrow \frac{\text{previous } \alpha_i \cdot loss_i}{\sum_{i=1}^n \alpha_i \cdot loss_i}$ 
8:   previous  $\alpha_i \leftarrow \alpha_i$ 
9:    $\alpha_i \leftarrow \alpha'_i$ 
10: end for
```

Due to limited time, we haven't fully explore it yet. Thus we simply include it in Appendix here.

B. Disentangle the sub-metric of Image Quality

As discussed in Section II-A, there could be interactions between different metrics. This observation leads to an intriguing question: is it possible to disentangle a parent metric into several child metrics? For instance, consider image quality. Despite numerous exceptional studies [5][6] that concentrate on enhancing image quality, to the best of my knowledge, the disentanglement of image quality into subsidiary metrics remains largely unexplored. Consequently, we propose the following approach: We begin by identifying some crucial metrics pertinent to image quality.

Firstly, we enumerate several significant metrics that are pertinent to image quality.

- **Resolution:** This is a basic measure of image clarity. A higher resolution entails richer details and more natural color gradients in the image.
- **Vivid Details:** This refers to whether the elements in the image are clearly depicted without any blurring or distortion. This includes whether the edges of the image are sharp, and whether textures, shadows, and highlights are clearly visible.
- **No Blur:** This refers to whether the image is clear, without any blurring or distortion. This could be due to the camera's focus being off, or blurring occurring in the image creation process.
- **Color Accuracy:** This refers to whether the colors in the image truthfully reflect the original scene. Color accuracy can be influenced by various factors, including lighting conditions, camera settings, and image editing.
- **Contrast:** This refers to the degree of difference between the bright and dark parts of an image. Good contrast can help to highlight the details and depth of the image.

TABLE VI

Results of Different Initialization. To prove the excellent performance of prompt design method is robust, we test it on three randomly picked different seed. We could see that there is no obvious performance gap between different seed, which is used for model parameters initialization and dataset shuffle.

| Task | Seed = 42 | | Seed = 100 | | Seed = 200 | |
|------------------|-----------|--------|------------|--------|------------|--------|
| | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC |
| AGIQA-3K Quality | 0.9694 | 0.9329 | 0.9530 | 0.9119 | 0.9577 | 0.9175 |
| AGIQA-3K Align | 0.9695 | 0.9392 | 0.9631 | 0.9195 | 0.9714 | 0.9363 |

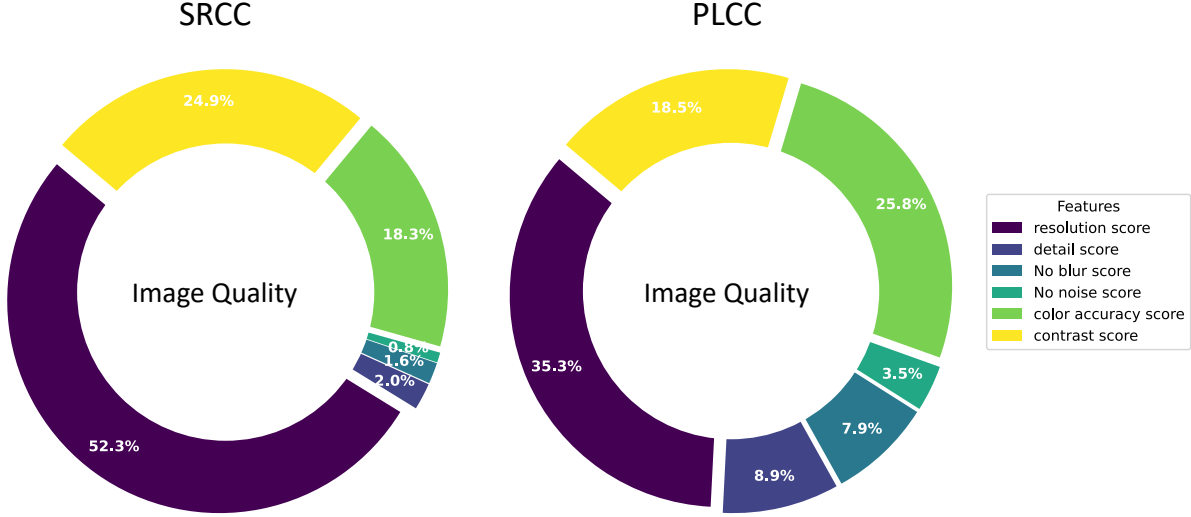


Fig. 6. Pie Chart Showing the Corresponding Ratio of Different Child Metrics. As we could see from the chart, there truly exists some difference between different child metrics, which could be explored further.

- **Noise Level:** This refers to random variations in color in the image, usually appearing as grainy or speckled discoloration. High-quality images should minimize noise as much as possible.

Next, we consider the question of quantification: How can we measure this pattern? The methodology we've developed employs specifically designed prompts. We utilize the score obtained from the prompt "extremely high quality image, with vivid details" as the base score, denoted as S_{base} . The scores from other prompts - such as "high resolution," "with vivid details," "without blur," "without noise," "with accurate color," and "with high contrast" - are denoted as S_i . We then calculate their relative significance in assessing image quality by determining their counting ratio:

$$ratio_i = \frac{1/(S_{base} - S_i)}{\sum_{i=1}^n \frac{1}{S_{base} - S_i}} \quad (4)$$

In the subsequent experiment, we utilized a model that was specifically trained to assess image quality using the prompt "high quality image." The results showed that $S_{base} > S_i$ for $i = 1, 2, \dots, n$, which aligns with our expectations. We've illustrated these findings in Figure 6. However, it's important to note that this approach is a preliminary trial and lacks robust theories. Recognizing its potential as an interesting research avenue, we have

included it in the Appendix for further exploration and discussion.

Acknowledgment

The authors would like to thank the Teacher of and T.A of CS3324 for providing valuable advice which greatly help me complete this assignment.

References

- [1] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong, "Imagereward: Learning and evaluating human preferences for text-to-image generation," 2023.
- [2] C. Li, Z. Zhang, H. Wu, W. Sun, X. Min, X. Liu, G. Zhai, and W. Lin, "Agiqa-3k: An open database for ai-generated image quality assessment," IEEE Transactions on Circuits and Systems for Video Technology, pp. 1–1, 2023.
- [3] J. Wang, H. Duan, J. Liu, S. Chen, X. Min, and G. Zhai, "Aigcqa2023: A large-scale image quality assessment database for ai generated images: from the perspectives of quality, authenticity and correspondence," 2023.
- [4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.
- [5] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 8110–8119.
- [6] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," IEEE transactions on pattern analysis and machine intelligence, vol. 44, no. 5, pp. 2567–2581, 2020.



Benhao Huang is presently in his third year of undergraduate studies at Shanghai Jiao Tong University, majored in Computer Science. He holds a research internship position at the SJTU Interpretable ML Lab, where he works under the guidance of Prof. Quanshi Zhang. Recently, his research interests have been centered around Interpretable AI, Understanding Human Activity, Visual Reasoning, and Graph Learning. You can visit his website at huskydoge.github.io for

more information. For any communication, he can be reached at hbh001098hbh@sjtu.edu.cn.