Exploring CLIP for Assessing the Look and Feel of Images

Jianyi Wang, Kelvin C.K. Chan, Chen Change Loy*

S-Lab, Nanyang Technological University
{jianyi001, chan0899, ccloy}@ntu.edu.sg

Abstract

Measuring the perception of visual content is a long-standing problem in computer vision. Many mathematical models have been developed to evaluate the *look* or quality of an image. Despite the effectiveness of such tools in quantifying degradations such as noise and blurriness levels, such quantification is loosely coupled with human language. When it comes to more abstract perception about the *feel* of visual content, existing methods can only rely on supervised models that are explicitly trained with labeled data collected via laborious user study. In this paper, we go beyond the conventional paradigms by exploring the rich visual language prior encapsulated in Contrastive Language-Image Pre-training (CLIP) models for assessing both the quality perception (*look*) and abstract perception (*feel*) of images without explicit task-specific training. In particular, we discuss effective prompt designs and show an antonym prompt pairing strategy to harness the prior. We also provide extensive experiments on controlled datasets and Image Quality Assessment (IQA) benchmarks. Our results show that CLIP captures meaningful priors that generalize well to different perceptual assessments.

Introduction

The *look* and *feel* are two contributing factors when humans interpret an image, and the understanding of these two elements has been a long-standing problem in computer vision. The *look* of an image is often related to quantifiable attributes that directly affect the content delivery, such as exposure and noise level. In contrast, the *feel* of an image is an abstract concept immaterial to the content and cannot be easily quantified, such as emotion and aesthetics. It is of common interest to explore the possibility of a universal understanding of both *look* and *feel*, as it can not only save efforts on manual labeling but also facilitate the development of vision tasks such as restoration (Zhang et al. 2019; Jo, Yang, and Kim 2020; Wenlong et al. 2021; Hui et al. 2021).

Considerable efforts have been devoted to the assessment of both the quality perception (*i.e.*, look) and the abstract perception (*i.e.*, feel) of images[1]. Earlier methods (Mittal, Moorthy, and Bovik 2012; Saad, Bovik, and Charrier 2012; Mittal, Soundararajan, and Bovik 2012; Zhang, Zhang, and Bovik 2015; Ma et al. 2017a) design image-level features, such as natural scene statistics (Ruderman 1994), for quality assessment. Despite the success of these methods, the optimality of hand-crafted features is often in doubt, and the correlation to human perception is inferior in general. To alleviate the problem, later methods (Ma et al. 2017b; Ke et al. 2021; Kong et al. 2016; Sheng et al. 2018; Jin et al. 2022;

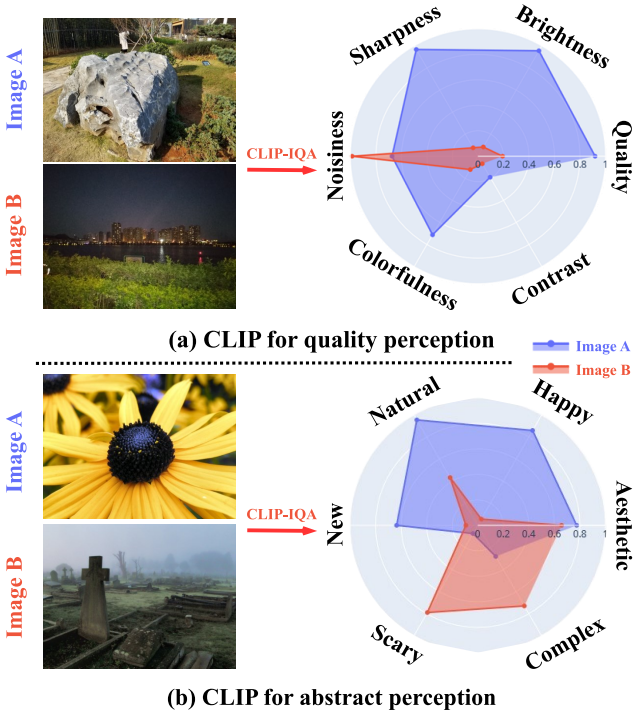


Figure 1: We explore the potential of CLIP for visual perception. We find that CLIP is capable of assessing quality (*e.g.*, "overall quality", "brightness", etc) as shown in (a). CLIP can be extended to the challenging task of abstract perception (*e.g.*, "aesthetic", "happy", etc) in (b). The assessment is done without explicit task-specific training using the proposed antonym prompts to harness prior in CLIP.

*Corresponding author.

[1]The notions of quality perception and abstract perception are used in this paper mainly to facilitate our analysis.

Kim et al. 2018; Yao et al. 2020; Achlioptas et al. 2021) resort to the learning-based paradigm, where a quality prediction model is directly learned from manually labeled data. With the presence of labeled data, these approaches demonstrate a much higher correlation to human perception. However, the laborious data labeling process and the task-specific nature of learning-based methods limit their generalizability to unseen attributes. More importantly, the aforementioned approaches are intended for either quality or abstract perception, possessing limited versatility[2].

We are interested in knowing whether there exists a notion that (1) does not require hand-crafted features and task-specific training, (2) possesses a high correlation to human perception, and (3) can handle both quality and abstract perception. To this end, we turn our focus to CLIP (Radford et al. 2021), a contrastive-learning-based visual-language pretraining approach. Through training with massive image-text pairs, CLIP demonstrates exceptional capability in building semantic relationship between texts and visual entities without explicit training (Patashnik et al. 2021; Jain, Tancik, and Abbeel 2021; Hessel et al. 2021; Rao et al. 2022; Zhou, Loy, and Dai 2022; Gu et al. 2022). Inspired by the appealing property of CLIP, we hypothesize that CLIP could have captured the relationship between human language and visual perception for image assessment.

The problem of harnessing CLIP for perception assessment can be more challenging compared to existing works related to objective attributes, such as image manipulation (Patashnik et al. 2021; Gabbay, Cohen, and Hoshen 2021; Xu et al. 2022), object detection (Gu et al. 2022; Zhong et al. 2022; Shi et al. 2022), and semantic segmentation (Rao et al. 2022; Zhou, Loy, and Dai 2022). Specifically, CLIP is known to be sensitive to the choices of prompts (Radford et al. 2021), and perception is an abstract concept with no standardized adjectives, especially for the *feel* of images. In addition, linguistic ambiguity (Khurana et al. 2017) (*e.g.*, "a clean image" can either refer to an image without noise or an image related to the action of cleaning) could severely affect perception assessment. As a result, the performance of CLIP on this task can be highly volatile, resulting from different choices of prompts.

Our study represents the first attempt to investigate the potential of CLIP on the challenging yet meaningful task of perception assessment. We begin our exploration by delving into the selection of prompts so that potential vagueness due to linguistic ambiguity can be minimized. To this end, we introduce a prompt pairing strategy where antonym prompts are adopted in pairs (*e.g.*, "Good photo." and "Bad photo."). Based on our strategy, we show that CLIP can be directly applied to visual perception assessment without any task-specific finetuning. For quality perception, we demonstrate that CLIP is able to assess the overall quality of an image by simply using "good" and "bad" as prompts, and achieve a high correlation to human's perception in common IQA datasets (Ghadiyaram and Bovik 2015; Hosu et al. 2020; Fang et al. 2020). Furthermore, we investigate the capability of CLIP in assessing fine-grained quality, such as brightness and noisiness. We apply CLIP on common restoration benchmarks (Wei et al. 2018; Bychkovsky et al. 2011; Xu et al. 2018; Rim et al. 2020) and synthetic data using fine-grained attributes, and it is shown that CLIP is capable of determining the fine-grained quality of an image (Fig. 1-(a)). In addition, the versatility of CLIP can be seen from its extension to abstract perception. In particular, CLIP also possesses superior performance when the quality-related prompts are replaced with abstract attributes (*e.g.*, "happy" and "sad"). Our results on both aesthetic benchmark (Murray, Marchesotti, and Perronnin 2012) and corresponding user studies show that CLIP succeeds in distinguishing images with different *feelings* following human perception (Fig. 1-(b)).

Given the success of CLIP in a wide range of vision tasks, we believe our exploration is timely. A powerful and versatile method for image assessment is indispensable. As the first work in this direction, we begin with a direct adaptation of CLIP to image assessment with a carefully designed prompt pairing strategy (named CLIP-IQA), followed by extensive experiments to examine the capability boundary of CLIP. We show that CLIP is able to assess not only the *look* but also the *feel* of an image to a satisfactory extent. We further analyze and discuss the limitations of CLIP on the task to inspire future studies.

CLIP for Visual Perception

Extending CLIP for Visual Perception

Antonym prompt pairing. As depicted in Fig. 2-(a), a straightforward way to exploit CLIP for perception assessment is to directly calculate the cosine similarity between the feature representations of a given prompt (*e.g.*, "Good photo") and a given image. Specifically, let $\boldsymbol{x} \in \mathbb{R}^C$ and $\boldsymbol{t} \in \mathbb{R}^C$ be the features (in vector form) from the image and prompt, respectively, where C is the number of channels. The final predicted score $s \in [0, 1]$ is computed as:

$$s = \frac{\boldsymbol{x} \odot \boldsymbol{t}}{||\boldsymbol{x}|| \cdot ||\boldsymbol{t}||}, \tag{1}$$

where $\odot$ denotes the vector dot-product and $||\cdot||$ represents the ℓ_2 norm.

While this naïve adoption achieves a huge success in existing literature (Patashnik et al. 2021; Hessel et al. 2021; Rao et al. 2022; Zhou, Loy, and Dai 2022), it is not viable for perception assessment, due to linguistic ambiguity (Khurana et al. 2017). Particularly, "a rich image" can either refer to an image with rich content or an image related to wealth. As shown in Fig. 2-(c), using CLIP with a single prompt shows poor correlation with human perception on common IQA datasets (Ghadiyaram and Bovik 2015; Hosu et al. 2020).

To address this problem, we propose a simple yet effective prompt pairing strategy. In order to reduce ambiguity, we adopt antonym prompts (*e.g.*, "Good photo." and "Bad photo.") as a pair for each prediction. Let $\boldsymbol{t}_1$ and $\boldsymbol{t}_2$ be the features from the two prompts opposing in meanings, we first compute the cosine similarity

$$s_i = \frac{\boldsymbol{x} \odot \boldsymbol{t}_i}{||\boldsymbol{x}|| \cdot ||\boldsymbol{t}_i||}, \quad i \in \{1, 2\}, \tag{2}$$

[2]More discussion of related work is included in the supplementary material.

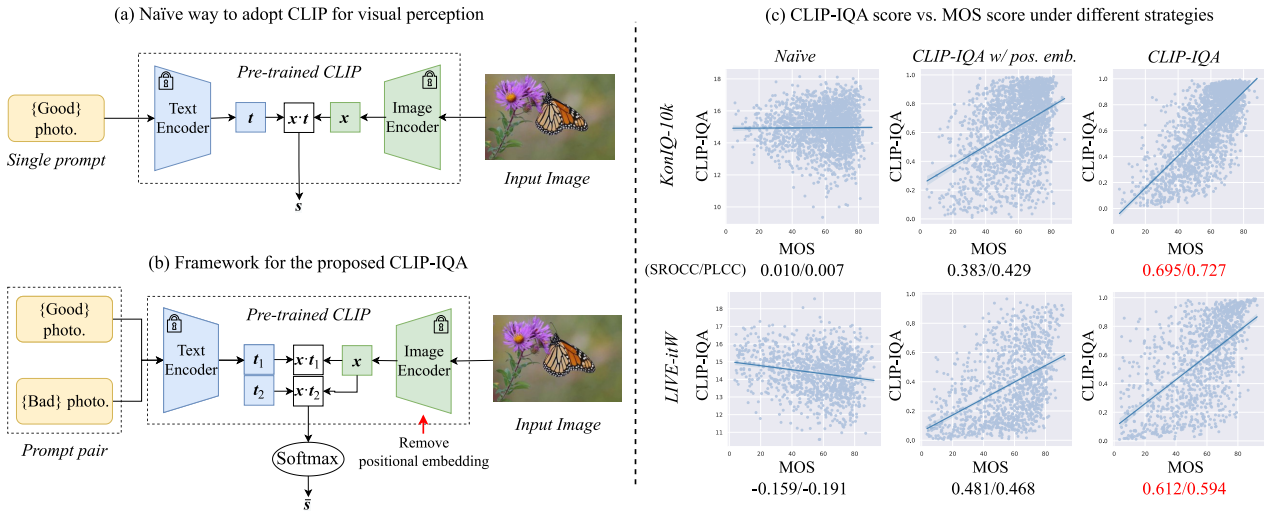


Figure 2: (a) A naïve approach of using a single prompt. (b) Our approach with (1) antonym prompt pairing strategy and (2) positional embedding removed. (c) Our method obtains much higher correlations with MOS scores in both KonIQ-10k (Hosu et al. 2020) and LIVE-itW (Ghadiyaram and Bovik 2015). The numbers in (c) represent SROCC/PLCC (higher is better).

and Softmax is used to compute the final score $\bar{s} \in [0, 1]$:

$$\bar{s} = \frac{e^{s_1}}{e^{s_1} + e^{s_2}}. \tag{3}$$

When a pair of adjectives is used, the ambiguity of one prompt is reduced by its antonym as the task is now cast as a binary classification, where the final score is regarded as a relative similarity. In particular, a larger $\bar{s}$ indicates a closer match to the corresponding attribute of $\boldsymbol{t}_1$. As shown in Fig. 2-(c), our proposed antonym prompt pairing significantly boosts the performance of CLIP – our method can predict scores more consistent with human-labeled MOS scores, reflecting from the higher Spearman's Rank-order Correlation Coefficient (SROCC) and Pearson's Linear Correlation Coefficient (PLCC).

Removal of positional embedding. Another limitation of CLIP is the requirement of fix-sized inputs. In particular, the ResNet-50-based CLIP requires an image with a size of 224×224. Such a requirement is unfavorable in perception assessment as the resizing and cropping operations may introduce additional distortions to the input image, altering the final score. For instance, the resizing operation leads to insensitivity to image resolution, and hence assessments related to resolution could be futile.

The above limitation results from the learnable positional embedding, which is fixed in size upon construction. Different from existing studies (Patashnik et al. 2021; Hessel et al. 2021; Jain, Tancik, and Abbeel 2021; Gu et al. 2022) that extensively use positional embedding, we conjecture that the positional embedding has a minimal effect on perception assessment as the primary focus of this task is to capture the perceptual relationship between an image and a given description. Therefore, we propose to remove the positional embedding to relax the size constraint. We adopt the ResNet variant since it can provide a stronger inductive bias to complement the removal of positional information. As shown in Fig. 2-(c), this relaxation further improves the correlation with human perception. We term our model **CLIP-IQA**, and all our experiments follow this setting.

Quality Perception

This section focuses on exploring the potential of CLIP-IQA on quality perception assessment. Specifically, we investigate its effectiveness on assessing the overall quality of images using conventional No-Reference IQA (NR-IQA) datasets (Ghadiyaram and Bovik 2015; Hosu et al. 2020; Fang et al. 2020). We then extend our scope to fine-grained attributes using synthetic data and common restoration benchmarks (Wei et al. 2018; Bychkovsky et al. 2011; Xu et al. 2018; Rim et al. 2020).

CLIP-IQA for overall quality. To assess the overall quality, we simply use one of the most commonly seen antonyms

["`Good photo.`", "`Bad photo.`"]

as the paired prompt for CLIP-IQA. We conduct experiments on three widely used NR-IQA benchmarks including LIVE-itW (Ghadiyaram and Bovik 2015) and KonIQ-10k (Hosu et al. 2020) for realistic camera distortions and SPAQ (Fang et al. 2020) for smartphone photography. We compare CLIP-IQA with nine NR-IQA methods, including four non-learning-based methods: BIQI (Moorthy and Bovik 2010), BLIINDS-II (Saad, Bovik, and Charrier 2012), BRISQUE (Mittal, Moorthy, and Bovik 2012), NIQE (Mittal, Soundararajan, and Bovik 2012), and five learning-based methods: CNNIQA (Kang et al. 2014), KonCept512 (Hosu et al. 2020), HyperIQA (Su et al. 2020), MUSIQ (Ke et al. 2021), VCRNet (Pan et al. 2022). For the non-learning-based methods, we take the numbers from (Hosu et al. 2020)

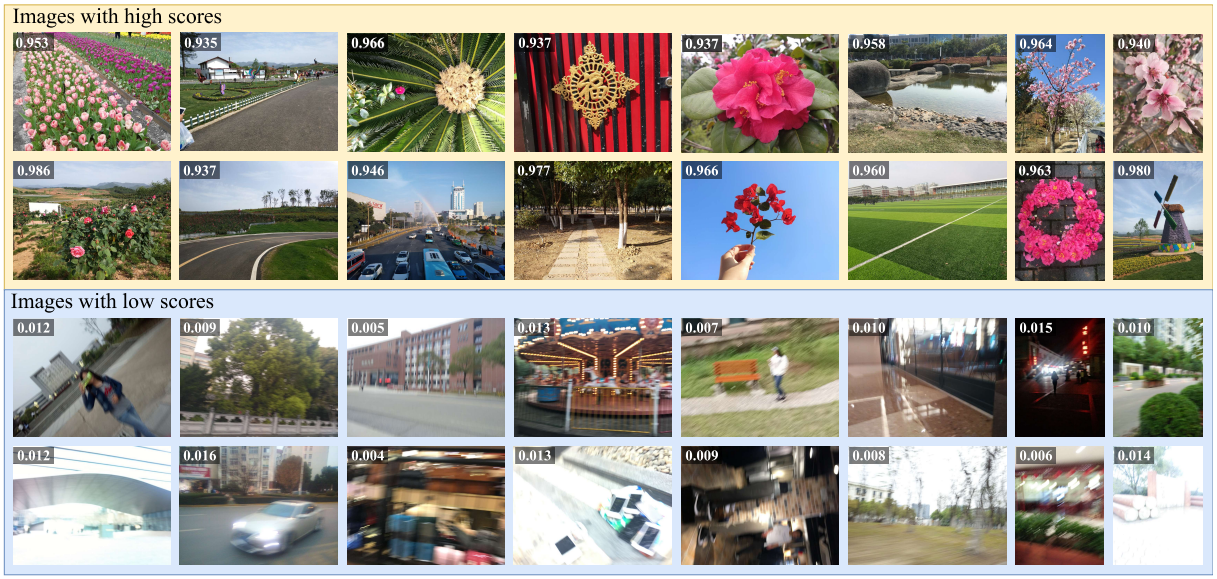


Figure 3: Best-scored and worst-scored images labeled by CLIP-IQA in SPAQ (Fang et al. 2020). CLIP-IQA performs well on SPAQ, i.e., assigning high scores to images with clear details and low scores to images with blur and low contrast. The numbers at the top left corner of each image indicate the CLIP-IQA score. (Zoom-in for best view)

Type	Methods	KonIQ-10k		LIVE-itW		SPAQ	
		SROCC↑	PLCC↑	SROCC↑	PLCC↑	SROCC↑	PLCC↑
w/o task-specific training	BIQI	0.559	0.616	0.364	0.447	0.591	0.549
	BLIINDS-II	0.585	0.598	0.090	0.107	0.317	0.326
	BRISQUE	0.705	0.707	0.561	0.598	0.484	0.481
	NIQE	0.551	0.488	0.463	0.491	0.703	0.670
	CLIP-IQA	0.695	0.727	0.612	0.594	0.738	0.735
w/ task-specific training	CNNIQA	0.572	0.584	0.465	0.450	0.664	0.664
	KonCept512	0.921	0.937	0.825	0.848	0.837	0.815
	HyperIQA	0.904	0.915	0.760	0.776	0.811	0.805
	MUSIQ	0.924	0.937	0.793	0.832	0.873	0.868
	VCRNet	0.894	0.909	0.678	0.701	0.781	0.766
	CLIP-IQA+	0.895	0.909	0.805	0.832	0.864	0.866

Table 1: Comparison to NR-IQA methods. Learning-based methods are trained on KonIQ-10k (Hosu et al. 2020) and tested on KonIQ-10k, LIVE-itW (Ghadiyaram and Bovik 2015) and SPAQ (Fang et al. 2020).

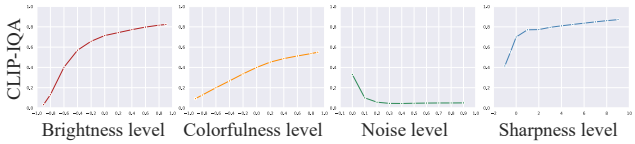


Figure 4: CLIP-IQA for fine-grained attributes on synthetic data with different input scales. CLIP-IQA clearly demonstrates positive correlations to the change of attributes.

for LIVE-itW and KonIQ-10k and adopt the official code for SPAQ. For learning-based methods except for MUSIQ[3], following Hosu *et al.* (Hosu et al. 2020), we train the models on KonIQ-10k dataset and test on the three datasets to assess their generalizability. Performance is evaluated with SROCC and PLCC. Detailed settings are discussed in the supplementary material.

As shown in Table 1, without the need of hand-crafted features, CLIP-IQA is comparable to BRISQUE and surpasses all other non-learning methods on all three benchmarks. In addition, even without task-specific training, CLIP-IQA outperforms CNNIQA, which requires training with annotations. The surprisingly good performance of CLIP-IQA verifies its potential in the task of NR-IQA. In Fig. 3 we show two sets of images obtaining high scores and low scores from CLIP-IQA respectively. It is observed that CLIP-IQA is able to distinguish images with different qualities.

In scenarios where annotations are available, CLIP-IQA can also be finetuned for enhanced performance. In this

[3]Since the training code is unavaliable, we adopt the official model trained on KonIQ-10k.

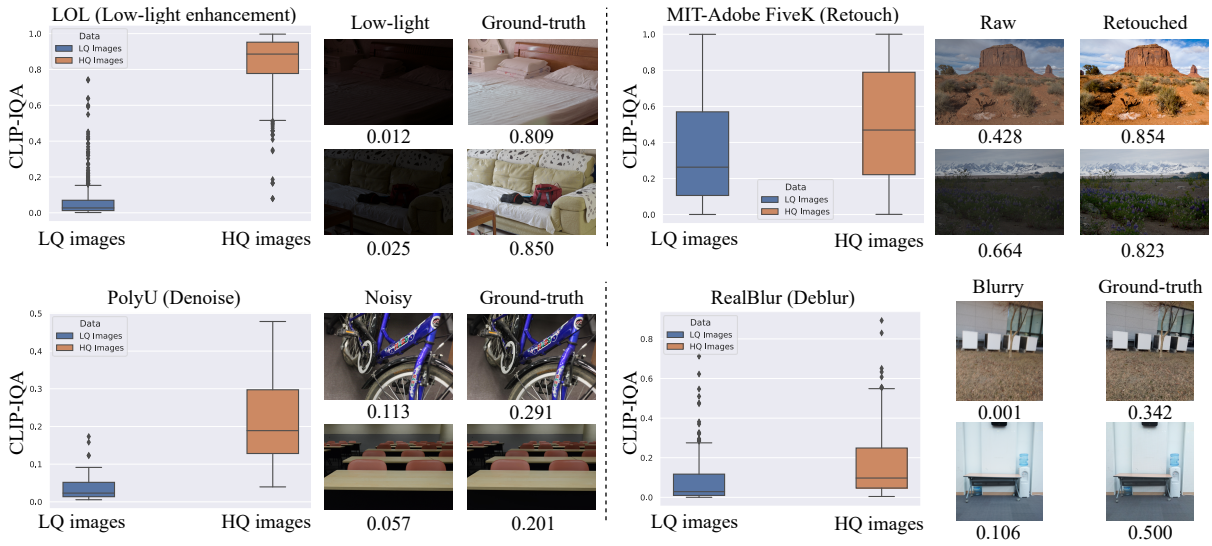


Figure 5: CLIP-IQA for fine-grained attributes on real-world benchmarks. Low-quality images receive a lower score from CLIP-IQA. The numbers under each image are the corresponding score from CLIP-IQA.

work, we develop a simple extension of CLIP-IQA, named CLIP-IQA$^+$, by using CoOp (Zhou et al. 2022) to fine-tune the prompts. Specifically, the prompts are initialized as ["`Good photo.`", "`Bad photo.`"] and updated with standard backpropagation[4]. The network weights are kept fixed. From Table 1 we see that CLIP-IQA$^+$ achieves SROCC/PLCC of 0.895/0.909, 0.805/0.832 and 0.864/0.866 on KonIQ-10k, LIVE-itW and SPAQ, respectively. These results are comparable to state-of-the-art deep learning based methods (Hosu et al. 2020; Su et al. 2020; Ke et al. 2021; Pan et al. 2022), showing the potential of large-scale language-vision training. Furthermore, it is observed that CLIP-IQA$^+$ possesses better generalizability. Specifically, while most learning-based methods experience significant performance drops on LIVE-itW and SPAQ, the drop of CLIP-IQA$^+$ is more subtle, and CLIP-IQA$^+$ achieves the second-best performance without task-specific architecture designs. We believe that the performance of CLIP-IQA$^+$ can be further improved by replacing CoOp with more sophisticated fine-tuning schemes such as Tip-Adaptor (Zhang et al. 2021). In addition to the generalizability, another advantage of CLIP-IQA$^+$ is the low storage cost for domain transfer. Unlike existing NR-IQA methods (*e.g.*, MUSIQ) that train different models for different domains (*e.g.*, technical quality, aesthetics quality), CLIP-IQA$^+$ needs to store only two prompts for each domain, significantly reducing the storage cost.

CLIP-IQA for fine-grained quality. Instead of assigning a single score to an image, humans usually judge an image from multiple perspectives, such as brightness, noisiness, and sharpness. Given the success of CLIP-IQA on the overall quality, we are curious about whether it can go beyond a single score and assess an image in a fine-grained fashion. To test the capability, we simply replace "good" and "bad" with the attribute we want to assess and its antonym, respectively. For example, we use "`Bright photo.`" and "`Dark photo.`" as the prompt when evaluating the brightness of images. It is noteworthy that unlike most learning-based approaches, CLIP-IQA does not require quality annotations. Therefore, CLIP-IQA is not confined to annotated labels and is able to assess arbitrary attributes required by users.

We first conduct our experiments on synthetic data for four representative attributes: *brightness*, *noisiness*, *colorfulness*, and *sharpness*. We adjust the extent of each attribute using the Python package `PIL`, and evaluate the performance on 200 images randomly selected from the test set of KonIQ-10k. From Fig. 4 we see that CLIP-IQA exhibits a high correlation to the changes of the attributes. More details are included in the supplementary material.

To further confirm its real-world applicability, we apply CLIP-IQA on four non-synthetic benchmarks: *LOL* (Wei et al. 2018) for low-light enhancement, *PolyU* (Xu et al. 2018) for denoising, *MIT-Adobe FiveK* (Bychkovsky et al. 2011) for retouching, and *RealBlur* (Rim et al. 2020) for deblurring. For each dataset, we apply CLIP-IQA to the low-quality images and high-quality images (*i.e.*, ground-truths) independently, and compare their quality scores. As depicted in Fig. 5, the quality scores of high-quality images are clearly above the low-quality ones, indicating that CLIP-IQA is able to identify fine-grained qualities.

Abstract Perception

While emotions and aesthetics are readily comprehensible by humans, the understanding of such concepts remains

[4]We adopt SGD with a learning rate of 0.002 during training. The model is trained for 100000 iterations with batch size 64 on KonIQ-10k dataset and the MSE loss is used to measure the distance between the predictions and the labels.

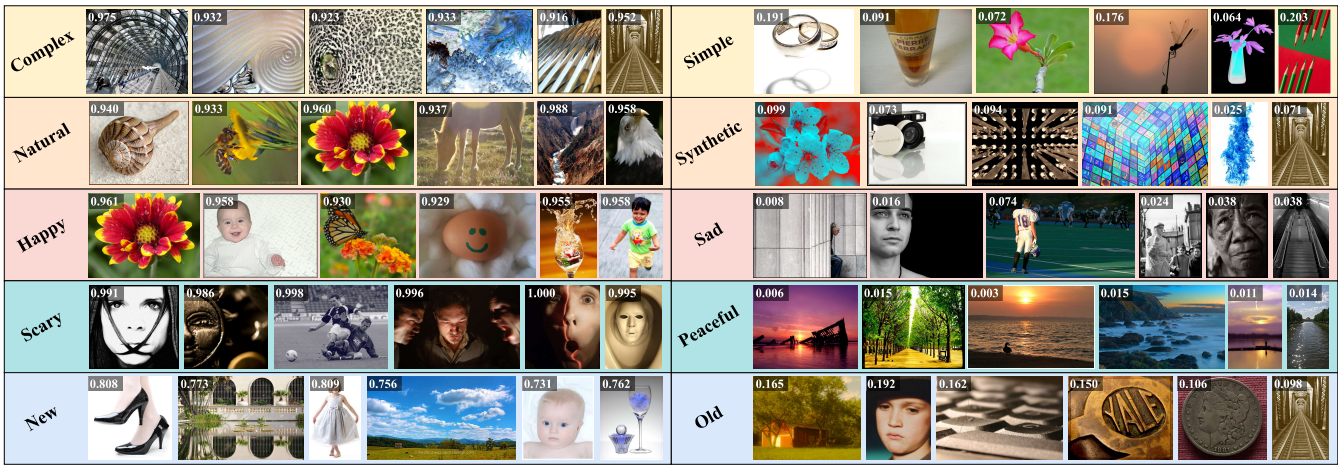
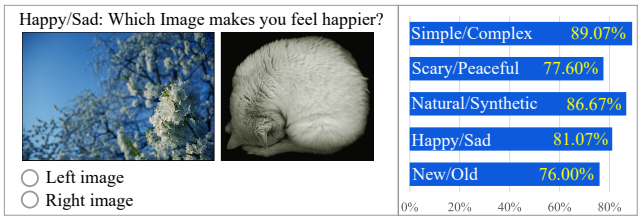


Figure 6: CLIP-IQA for assessing abstract perception. The results on these five attributes show that CLIP-IQA is able to understand abstract perception, *e.g.* *"Complex/Simple"*. The left and right are two sets of images selected according to the CLIP-IQA scores (numbers at the top left corner of each image) on abstract-attribute antonym pairs. (Zoom-in for best view)



(a) User study settings (b) Accuracy of user study

Figure 7: User study for abstract perception. (a) We ask the subjects to compare the two images and select the one more consistent with the given description. (b) The prediction of CLIP-IQA is consistent with human about 80% of the time.

non-trivial for machines. This section investigates the competence of CLIP-IQA in interpreting *abstract perception*. We conduct our experiments on the AVA dataset (Murray, Marchesotti, and Perronnin 2012) since it is among the largest visual perception datasets, containing over 250,000 images with a broad variety of content.

We evaluate the performance of CLIP-IQA using five attributes related to human emotions and artistic sensations, *i.e.*, *complex/simple*, *natural/synthetic*, *happy/sad*, *scary/peaceful*, and *new/old*. For each attribute, we compute the score for each image, and images are sorted according to their scores. In Fig. 6 we show the best-scored and worst-scored images for each attribute. It is observed that CLIP-IQA is able to perceive the abstract aspects of an image. Similar to quality attributes, CLIP-IQA is able to assess an image from different perspectives. For instance, the last image in the *complex* category is also *synthetic* and *old*. Such a fine-grained abstract understanding again demonstrates the effectiveness of the prior captured by CLIP during vision-language pretraining.

We further conduct a user study to verify the effectiveness of CLIP-IQA on abstract perception. The settings of the study are depicted in Fig. 7-(a). Specifically, we generate 15 image pairs for each of the above five attributes. For each pair, we label the one with higher score as positive (*e.g.*, happy) and lower score as negative (*e.g.*, sad). Each pair is rated by 25 subjects. Subjects are asked to compare the two images and choose the one more consistent with the provided description (*e.g.*, In the two images, which one makes you feel happier?). We then compute the classification accuracy of CLIP-IQA using the subject scores as ground-truths. As depicted in Fig. 7, CLIP-IQA achieves an accuracy of about 80% on all five attributes, indicating the great potential of CLIP-IQA on abstract idea understanding.

Discussion

Prompt Designs

While it is verified in previous works (Radford et al. 2021; Zhou et al. 2022; Rao et al. 2022) that the selection of prompts could impact the performance, such conclusion has not been drawn in the task of perception assessment. Therefore, we conduct empirical studies to investigate the effects brought by different choices of prompts.

First, it is observed that the choice of prompt templates has a significant effect to CLIP-IQA. We compare the performance of three different choices of templates from existing literature (Rao et al. 2022; Zhou, Loy, and Dai 2022) – (1) "`[text] photo.`", (2) "`A photo of [text].`", and (3) "`There is [text] in the photo.`". As shown in Table 2, noticeable differences are observed when using different templates. In this work, we adopt "`[text] photo.`" for simplicity, and we believe that there exist more sophisticated templates that lead to enhanced performance. Such exploration is left as our future work.

Next, we investigate the influence of adjectives with the above template. Similarly, the performance varies with

Settings				KonIQ-10k		LIVE-itW	
Template	Adjective	Backbone	Pos. Embedding	SROCC	PLCC	SROCC	PLCC
(1)	"Good/Bad"	ResNet-50	✗	0.695	0.727	0.612	0.594
(2)		ResNet-50	✗	0.116	0.119	0.263	0.276
(3)		ResNet-50	✗	0.214	0.217	0.347	0.351
(1)	"High quality/Low quality"	ResNet-50	✗	0.537	0.570	0.462	0.429
	"High definition/Low definition"	ResNet-50	✗	0.592	0.580	0.611	0.560
(1)	"Good/Bad"	ResNet-50	Vanilla	0.383	0.429	0.481	0.400
(1)	"Good/Bad"	ResNet-50	Interpolated	0.682	0.690	0.583	0.555
(1)	"Good/Bad"	VIT-B/32	Vanilla	0.416	0.464	0.488	0.479
(1)	"Good/Bad"	VIT-B/32	Interpolated	0.634	0.643	0.503	0.491
(1)	"Good/Bad"	VIT-B/32	✗	0.391	0.374	0.375	0.365

Table 2: Comparison of CLIP-IQA variants. It is observed that the choices of prompts and backbones have a significant effect on the performance. The templates are (1) "`[text] photo.`", (2) "`A photo of [text].`", and (3) "`There is [text] in the photo`, respectively. The first row is the settings for CLIP-IQA.

the adjectives selected[5]. For instance, as shown in Table 2, when assessing the overall quality of an image, "`Good/Bad`" achieves a better correlation to human perception than "`High quality/Low quality`" and "`High definition/Low definition`". We conjecture that such a phenomenon indicates poorer performance of uncommon adjectives. We remark here that this challenge is especially notable in perception assessment due to the existence of synonyms. The sensitivity to the choices of prompts indicates the need of a more comprehensive understanding of prompt designs.

Backbone of Image Encoder

The backbone of the image encoder significantly affects the performance of CLIP-IQA. Unlike convolutional models that implicitly capture positional information (Xu et al. 2021), Transformer relies heavily on the positional embedding to provide such clue. Hence, the Transformer variant of CLIP incorporates the positional embedding in earlier layers to compensate for the lack of such inductive bias. In contrast, the ResNet variant depends less on the positional embedding, and adopts the embedding only in deep layers during multi-head attention. Therefore, it is expected that the Transformer variant would experience a larger performance drop when the positional embedding is removed.

To verify the hypothesis, we conduct experiments on KonIQ-10k and LIVE-itW. As shown in Table 2, while the Transformer variant shows better performance than the ResNet variant with the positional embedding, a significant drop is observed in the Transformer variant when the embedding is removed. This result corroborates our hypothesis that positional embedding is more crucial in Transformer.

We further compare the performance of the ResNet variants with (1) positional embedding removal, and (2) positional embedding interpolation. We notice from Table 2 that the performance of the ResNet variant with positional embedding removal is better than that using positional embedding interpolation. We conjecture this is due to the inaccurate positional embedding caused by interpolation. Given the importance of accepting arbitrary-sized inputs in perception assessment, we adopt the ResNet variant with positional embedding removed throughout our experiments.

[5]We collect adjectives from various image assessment benchmarks and photo sharing websites, which demonstrate decent results on KonIQ-10k and LIVE-itW.

Limitations

Despite the encouraging performance of CLIP-IQA, there are challenges that remain unresolved. First, CLIP-IQA is sensitive to the choices of prompts. Therefore, it is of great importance to develop more systematic selection of prompts. In the future, our attention will be devoted to the design of prompts for improved performance.

Second, through training with vision-language pairs, CLIP is able to comprehend words and phrases widely used in daily communication. However, it remains a non-trivial problem for CLIP to recognize professional terms that are relatively uncommon in human conversations, such as *"Long exposure"*, *"Rule of thirds"*, *"Shallow DOF"*. Nevertheless, we believe that this problem can be attenuated by pretraining CLIP with such pairs.

Third, although our exploration shows the capability of CLIP on versatile visual perception tasks without explicit task-specific training, there still exist performance gaps between CLIP-IQA and existing task-specific methods due to the lack of task-specific architecture designs in CLIP-IQA. We believe the fusion of task-specific designs and the vision-language prior would achieve promising performance and will be a new direction in the task of perception assessment.

Conclusion

The remarkable versatility of CLIP has aroused great interests from computer vision researchers. This paper diverges from existing works and investigates the effectiveness of CLIP on perceiving subjective attributes. From our exploration, we find that CLIP, when equipped with suitable modifications, is able to understand both quality and abstract perceptions of an image. By providing a solid ground for CLIP-based perception assessment, we believe our studies and discussion could motivate future development in various domains, such as sophisticated prompts, better generalizability, and effective adoption of CLIP prior.

Acknowledgements

This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-PhD-2022-01-033[T]), the RIE2020 Industry Alignment Fund Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s). It is also partially supported by the NTU NAP grant. We thank Kede Ma, Yuming Fang and Hanwei Zhu for providing valuable technical details of SPAQ dataset.

References

Achlioptas, P.; Ovsjanikov, M.; Haydarov, K.; Elhoseiny, M.; and Guibas, L. J. 2021. Artemis: Affective language for visual art. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bychkovsky, V.; Paris, S.; Chan, E.; and Durand, F. 2011. Learning Photographic Global Tonal Adjustment with a Database of Input / Output Image Pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Fang, Y.; Zhu, H.; Zeng, Y.; Ma, K.; and Wang, Z. 2020. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gabbay, A.; Cohen, N.; and Hoshen, Y. 2021. An image is worth more than a thousand words: Towards disentanglement in the wild. *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*.

Ghadiyaram, D.; and Bovik, A. C. 2015. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing (TIP)*.

Gu, X.; Lin, T.-Y.; Kuo, W.; and Cui, Y. 2022. Open-vocabulary Object Detection via Vision and Language Knowledge Distillation. *Proceedings of International Conference on Learning Representations (ICLR)*.

Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *The 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*.

Hosu, V.; Lin, H.; Sziranyi, T.; and Saupe, D. 2020. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing (TIP)*.

Hui, Z.; Li, J.; Wang, X.; and Gao, X. 2021. Learning the Non-differentiable Optimization for Blind Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jain, A.; Tancik, M.; and Abbeel, P. 2021. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jin, X.; Lou, H.; Heng, H.; Li, X.; Cui, S.; Zhang, X.; and Li, X. 2022. Pseudo-labelling and Meta Reweighting Learning for Image Aesthetic Quality Assessment. *arXiv preprint arXiv:2201.02714*.

Jo, Y.; Yang, S.; and Kim, S. J. 2020. Investigating loss functions for extreme super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (CVPR-W)*.

Kang, L.; Ye, P.; Li, Y.; and Doermann, D. 2014. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; and Yang, F. 2021. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Khurana, D.; Koli, A.; Khatter, K.; and Singh, S. 2017. Natural language processing: State of the art, current trends and challenges. *arXiv preprint arXiv:1708.05148*.

Kim, H.-R.; Kim, Y.-S.; Kim, S. J.; and Lee, I.-K. 2018. Building emotional machines: Recognizing image emotions through deep neural networks. *IEEE Transactions on Multimedia (TMM)*.

Kong, S.; Shen, X.; Lin, Z.; Mech, R.; and Fowlkes, C. 2016. Photo Aesthetics Ranking Network with Attributes and Content Adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Ma, C.; Yang, C.-Y.; Yang, X.; and Yang, M.-H. 2017a. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding (CVIU)*.

Ma, K.; Liu, W.; Zhang, K.; Duanmu, Z.; Wang, Z.; and Zuo, W. 2017b. End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing (TIP)*.

Mittal, A.; Moorthy, A. K.; and Bovik, A. C. 2012. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing (TIP)*.

Mittal, A.; Soundararajan, R.; and Bovik, A. C. 2012. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*.

Moorthy, A. K.; and Bovik, A. C. 2010. A two-step framework for constructing blind image quality indices. *IEEE Signal Processing Letters*.

Murray, N.; Marchesotti, L.; and Perronnin, F. 2012. AVA: A large-scale database for aesthetic visual analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pan, Z.; Yuan, F.; Lei, J.; Fang, Y.; Shao, X.; and Kwong, S. 2022. VCRNet: Visual Compensation Restoration Network for No-Reference Image Quality Assessment. *IEEE Transactions on Image Processing (TIP)*.

Patashnik, O.; Wu, Z.; Shechtman, E.; Cohen-Or, D.; and Lischinski, D. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of International Conference on Machine Learning (ICML)*.

Rao, Y.; Zhao, W.; Chen, G.; Tang, Y.; Zhu, Z.; Huang, G.; Zhou, J.; and Lu, J. 2022. DenseCLIP: Language-Guided Dense Prediction with Context-Aware Prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rim, J.; Lee, H.; Won, J.; and Cho, S. 2020. Real-World Blur Dataset for Learning and Benchmarking Deblurring Algorithms. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Ruderman, D. L. 1994. The statistics of natural images. *Network: Computation in Neural Systems*.

Saad, M. A.; Bovik, A. C.; and Charrier, C. 2012. Blind image quality assessment: A natural scene statistics approach in the DCT domain. *IEEE Transactions on Image Processing (TIP)*.

Sheng, K.; Dong, W.; Ma, C.; Mei, X.; Huang, F.; and Hu, B.-G. 2018. Attention-based multi-patch aggregation for image aesthetic assessment. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*.

Shi, H.; Hayat, M.; Wu, Y.; and Cai, J. 2022. ProposalCLIP: Unsupervised Open-Category Object Proposal Generation via Exploiting CLIP Cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Su, S.; Yan, Q.; Zhu, Y.; Zhang, C.; Ge, X.; Sun, J.; and Zhang, Y. 2020. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wei, C.; Wang, W.; Yang, W.; and Liu, J. 2018. Deep Retinex Decomposition for Low-Light Enhancement. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Wenlong, Z.; Yihao, L.; Dong, C.; and Qiao, Y. 2021. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Xu, J.; Li, H.; Liang, Z.; Zhang, D.; and Zhang, L. 2018. Real-world noisy image denoising: A new benchmark. *arXiv preprint arXiv:1804.02603*.

Xu, R.; Wang, X.; Chen, K.; Zhou, B.; and Loy, C. C. 2021. Positional encoding as spatial inductive bias in gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xu, Z.; Lin, T.; Tang, H.; Li, F.; He, D.; Sebe, N.; Timofte, R.; Van Gool, L.; and Ding, E. 2022. Predict, Prevent, and Evaluate: Disentangled Text-Driven Image Manipulation Empowered by Pre-Trained Vision-Language Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yao, X.; Zhao, S.; Lai, Y.-K.; She, D.; Liang, J.; and Yang, J. 2020. APSE: Attention-aware polarity-sensitive embedding for emotion-based image retrieval. *IEEE Transactions on Multimedia (TMM)*.

Zhang, L.; Zhang, L.; and Bovik, A. C. 2015. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing (TIP)*.

Zhang, R.; Fang, R.; Gao, P.; Zhang, W.; Li, K.; Dai, J.; Qiao, Y.; and Li, H. 2021. Tip-Adapter: Training-free CLIP-Adapter for Better Vision-Language Modeling. *arXiv preprint arXiv:2111.03930*.

Zhang, W.; Liu, Y.; Dong, C.; and Qiao, Y. 2019. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Zhong, Y.; Yang, J.; Zhang, P.; Li, C.; Codella, N.; Li, L. H.; Zhou, L.; Dai, X.; Yuan, L.; Li, Y.; et al. 2022. RegionCLIP: Region-based Language-Image Pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhou, C.; Loy, C. C.; and Dai, B. 2022. DenseCLIP: Extract Free Dense Labels from CLIP. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*.