Image Quality and Abstract Perception Evaluation

Eryn Ma
Carnegie Mellon University
Pittsburgh, PA 15213
erynm@andrew.cmu.edu

Hail Song
Carnegie Mellon University
Pittsburgh, PA 15213
hails@andrew.cmu.edu

Yuchen Dai
Carnegie Mellon University
Pittsburgh, PA 15213
yuchenda@andrew.cmu.edu

Shiwon Kim
Carnegie Mellon University
Pittsburgh, PA 15213
shiwonk@andrew.cmu.edu

Abstract

Human perception of images encompasses both objective qualities, such as technical image quality, and subjective abstract perceptions, including mood, emotion, and aesthetics. While traditional Image Quality Assessment (IQA) approaches rely on full-reference or no-reference models, they are limited in capturing intangible aesthetic qualities and require extensive labeled datasets. Vision-language models, such as CLIP and its derivatives, offer a promising alternative for bridging image content with human interpretation but still face challenges in optimizing for subjective evaluation. To address this gap, we propose CLIP-UIQA, a novel architecture combining CLIP-IQA and UIQA with a multi-branch backbone. This approach integrates global aesthetic characteristics, local distortions, and salient features to enhance the interpretation of abstract image perceptions. Leveraging curated datasets and metrics for both objective and perceptual quality assessment, our model demonstrates superior accuracy and faster convergence compared to existing methods. The results validate the potential of CLIP-UIQA in applications such as content recommendation, automated media curation, and user-centered media evaluations. Code available at: https://github.com/hailsong/CLIP-IQA

1 Introduction

1.1 Background and Motivation

Human perception of beauty and pleasure in images is deeply subjective, shaped by factors such as cultural context, personal experiences, and emotional states. Understanding this perception is challenging for computer vision researchers, especially as they aim to replicate both objective and subjective evaluations of images. Image Quality Assessment (IQA) and abstract perception are two fields that attempt to quantify human experience with images, with IQA focusing on measurable qualities like brightness and color, and abstract perception interpreting intangible aspects like mood and emotion.

1.2 Traditional Approaches in BIQA

Early BIQA methods often relied on Full-Reference (FR) and No-Reference (NR) models. FR models assess quality by comparing an image to a pristine reference, while NR models, especially Blind Image Quality Assessment (BIQA), assess images independently of reference data. Traditional NR-BIQA approaches have largely depended on hand-crafted features and large labeled datasets,

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

limiting flexibility and scalability. These approaches are particularly restricted in capturing subjective qualities, as they require extensive human annotation to cover diverse image content and distortion types.

1.3 Recent Advances: Vision-Language Models

Recent advancements in vision-language models, such as Contrastive Language-Image Pretraining (CLIP), offer a flexible approach to BIQA by using natural language prompts to assess images. CLIP-based approaches like CLIP-IQA guide models in evaluating both technical (e.g., brightness) and subjective (e.g., mood) aspects, aligning closely with human aesthetic judgments. This shift not only reduces the need for labeled data but also enables new applications in content recommendation, media curation, and personalized image assessments that adapt to user preferences.

1.4 Challenges and Contributions of This Work

Despite these advances, current BIQA models are still limited in their ability to fully capture the subjective, nuanced qualities that define human aesthetic experience. While vision-language models like Wang et al. [1] have achieved promising results in assessing both quality and abstract perception, there is still a need for more targeted optimization to sensitively evaluate the *feel* of an image, including aspects such as mood, emotion, and aesthetic appeal. This is especially relevant for applications where subtle and contextual image assessments are crucial, such as personalized media curation, where users may have unique preferences that are not easily captured by objective quality metrics alone.

In this work, we build upon existing BIQA models to address the challenge of abstract perception with greater sensitivity and depth. We contribute a new approach that enhances the model's ability to interpret intangible qualities by designing optimized prompts and employing a multitask learning framework. By fine-tuning these prompts and introducing auxiliary tasks, our model is designed to bridge the gap between human and machine perception, aiming to capture the complexity of human aesthetic judgments across diverse contexts. Ultimately, we anticipate that our approach will enrich applications that demand sophisticated and user-centered media evaluations, providing more nuanced insights for content recommendation, automated curation, and user-centric aesthetic assessments.

2 Literature Review

2.1 Blind Image Quality Assessment (BIQA)

Blind Image Quality Assessment (BIQA) aims to understand how humans perceive image quality without needing a reference image. Early BIQA methods fell under the No-Reference (NR) category, as they sought to measure image quality without requiring a pristine reference image. Traditional NR methods, such as those based on natural scene statistics (NSS), relied on hand-crafted features to capture measurable elements like texture and noise. While these methods provided some insight into quality, they struggled with generalizing to more complex, subjective aspects of image quality [1, 2].

With advancements in deep learning, NR-BIQA evolved through models trained on large labeled datasets, allowing them to learn quality indicators directly and align more closely with human judgments. Models like KonCept512 [3] and HyperIQA [4] are prominent examples of NR-BIQA models that leverage deep learning to achieve stronger generalization. These methods, however, rely heavily on labeled data, which is often gathered through labor-intensive Mean Opinion Scores (MOS). This dependency on human labeling not only increases cost but also limits the generalizability of NR-BIQA models to diverse image contents and distortion types [1]. Full-Reference (FR) IQA methods, while effective in controlled settings, are unsuitable for blind assessment tasks like BIQA since they require a reference image to compute quality scores. Therefore, NR methods remain the primary approach for blind image quality assessment.

Recently, vision-language models like CLIP have emerged as versatile tools for NR-BIQA, as they can learn correlations between images and natural language without the need for specialized labeled data. CLIP-IQA, for instance, utilizes antonym-based prompt pairs (e.g., "Good photo" vs. "Bad photo") to guide the model in assessing both objective qualities (like brightness) and abstract qualities (such as mood), making it highly adaptable to diverse assessments without additional labeled data [1].

Building on this vision-language paradigm, Zhang et al. [2] introduced the LIQE model, which combines NR-BIQA with multitask learning by integrating auxiliary tasks such as scene classification and distortion identification. This multitask model further improves BIQA performance by incorporating scene and distortion information, although it adds complexity due to the need for precise parameter sharing across tasks. While LIQE represents a merging of NR-BIQA with other vision tasks, the FR and NR approaches themselves remain distinct within the BIQA field, as NR continues to be essential for tasks where reference images are unavailable.

Benchmark datasets like Aesthetic Visual Analysis (AVA) [5] and KonIQ-10k [3] are essential for evaluating NR-BIQA models, as they contain real-world distortions and aesthetic ratings that challenge models to handle both objective and subjective aspects of image quality. By training on these datasets, models like CLIP-IQA and LIQE demonstrate the flexibility and adaptability of vision-language approaches, which excel in NR tasks without task-specific tuning [1, 2].

2.2 Abstract Perception

The field of abstract perception in computer vision remains relatively underexplored compared to Image Quality Assessment (IQA). While IQA encompasses both Full-Reference (FR) and No-Reference (NR) approaches for measurable qualities like brightness, sharpness, and noise, abstract perception goes beyond these metrics to assess subjective elements, such as emotion and aesthetics, that cannot be easily quantified. Vision-language models like CLIP have opened new possibilities for NR assessment in abstract perception by using natural language prompts to bridge image content and human interpretations. For instance, Wang et al. [1] introduced antonym-based prompt engineering to interpret abstract qualities through prompt-pairing techniques that align closely with human judgments. This approach leverages CLIP's pretraining, allowing it to interpret subjective aspects like mood and style without task-specific training or labeled data [1, 6].

While recent works like Zhang et al. [2] expand NR-BIQA through multitask models by integrating auxiliary tasks such as scene classification and distortion type identification, their modules are designed primarily for quality assessment rather than abstract perception. Nevertheless, CLIP-based models remain more adaptable for NR assessments of abstract qualities in images, such as aesthetic value or emotional tone, which are challenging for traditional models trained on content classification datasets like ImageNet [1, 6, 7]. The flexibility of vision-language models in handling diverse visual attributes without extensive labeled data makes them valuable tools for abstract perception tasks [8, 9].

3 Baseline Selection

3.1 Baseline Candidates

3.1.1 CLIP-IQA

CLIP-IQA, proposed by Wang et al. [1], outperforms existing image quality evaluation models by leveraging a pre-trained CLIP model, which was initially used in image diffusion models. By using CLIP's rich visual-language priors, their method can assess not only the quality perception (i.e., *look*) but also abstract perception (how people *feel* when they see an image). This ability is validated through user studies, confirming the model's alignment with human perception.

3.1.2 LIQE

Zhang et al. [2] built upon CLIP and developed one of the state-of-the-art No-Reference Image Quality Assessment (NR-IQA) model, Language-Image Quality Evaluator (LIQE), outperforming CLIP-IQA. Zhang et al. [2] clarified in their paper while they also mainly leveraged CLIP, their new contribution came from exploring CLIP in the multi-task learning setting to enhance auxiliary knowledge transfer. In addition, they fine-tuned the CLIP model instead of prompt tuning to reach better quality prediction performance.

3.2 Model Selection

In selecting a baseline model, we considered both the CLIP-IQA [1] and LIQE [2] due to their robust performance in the realm of Image Quality Assessment (IQA) and their grounding in vision-language

models. CLIP-IQA was designed to evaluate both IQA and abstract perception tasks, leveraging the pretrained CLIP model to assess objective image quality (e.g., brightness, noise level) as well as more abstract elements such as aesthetic appeal and emotional tone. This model's flexibility stems from its use of antonym-based prompt engineering, enabling it to effectively capture both concrete and abstract qualities without extensive task-specific training. On the other hand, LIQE enhances the IQA performance further by incorporating multitask learning with additional scene classification and distortion type identification modules, making it highly specialized for IQA tasks.

Despite the higher accuracy and advanced capabilities of LIQE in IQA, we chose CLIP-IQA as our baseline due to our project's focus on abstract perception rather than strict IQA. Unlike LIQE, which has specialized modules for scene and distortion classification [2], CLIP-IQA is structured to assess a wider range of perceptual attributes, allowing us to capture intangible qualities such as emotion and aesthetic *feel* more directly [1]. This makes CLIP-IQA a better fit for our project, as it provides a balanced approach to both objective and subjective image qualities, aligning well with our goal of understanding abstract perception in images. Their model is published on GitHub.

4 Baseline Model Description

4.1 CLIP

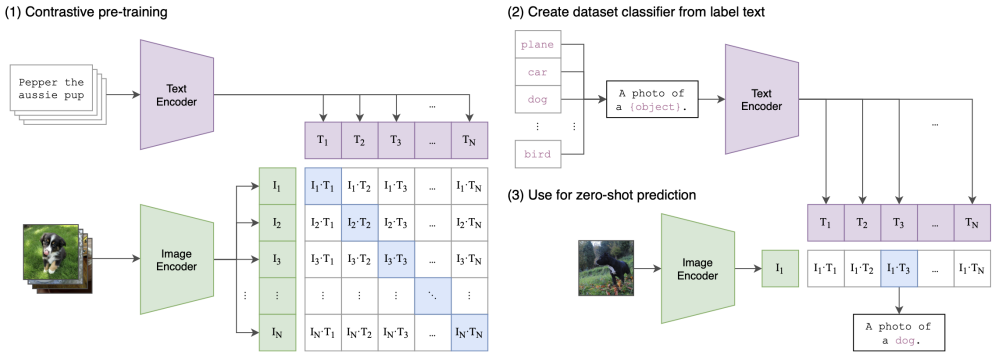

Figure 1: Summary of CLIP. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

The model aims to learn representations that align images with text descriptions in a shared embedding space.

For N image text pairs in a batch, $\mathbf{I}_i$ is the embedding of the i-th image and $\mathbf{T}_i$ is the embedding of the i-th text. τ is a learnable temperature parameter to make the softmax distribution steeper.

4.1.1 Contrastive Pre-training

In the first stage, as shown in Figure 1 (1),images and texts are tokenized through image encoder (ResNet or Vision Transformer) and text encoder (CBOW or Text Transformer). The texts and the images tokens then linearly projected a contrastive embedding.

The cosine similarity for visual and text embedding is calculated as such $\text{sim}(\mathbf{I}_i, \mathbf{T}_j) = \frac{\mathbf{I}_i^\top \mathbf{T}_j}{\|\mathbf{I}_i\| \|\mathbf{T}_j\|}$.

$\ell_i^{(I \to T)}$ and $\ell_i^{(T \to I)}$ represent the cross entropy of the softmax distribution of the row and the column of the contrastive embedding matrix respectively.

$$\ell_i^{(I \to T)} = -\log \frac{\exp\left(\langle \mathbf{I}_i, \mathbf{T}_i \rangle / \tau\right)}{\sum_{k=1}^{N} \exp\left(\langle \mathbf{I}_i, \mathbf{T}_k \rangle / \tau\right)} \tag{1}$$

$$\ell_i^{(T\to I)} = -\log \frac{\exp\left(\langle \mathbf{T}_i, \mathbf{I}_i \rangle / \tau\right)}{\sum_{k=1}^{N} \exp\left(\langle \mathbf{T}_i, \mathbf{I}_k \rangle / \tau\right)} \tag{2}$$

The model is then backpropagated to minimize the following objective function, in which the losses of image to text and text to image are weighted and averaged:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \left(\lambda\, \ell_i^{(I\to T)} + (1-\lambda)\, \ell_i^{(T\to I)} \right) \tag{3}$$

The objective maximizes the cosine similarity of correct image-text pairs while minimizing it for incorrect pairs by maximizing the inner products on the diagonal highlighted by blue and minimizing other white boxes.

4.1.2 Zero-Shot Prediction

In the second stage, as shown in (2) and (3) in Figure 1, a restricted set of categories (e.g., "photo", "car") is passed into the model to form a series of prompts (e.g., "a photo of a car"). The prompt is then passed through a text encoder to be linearly projected to the contrastive embedding space, where the target image will be passed in, through similar encoder and linear projection process, to generate the prediction. The prompt has been shown to greatly enhance the performance and the restricted set largely reducses the probability space the model needs to search for.

4.2 CLIP-IQA

While the cosine similarity is very successful, it does not perform very well in the face of linguistic ambiguity (e.g., "a rich image" could both mean an image with dense content or an image with reference to wealth). Wang et al. [1] proposed a novel framework, as shown in Figure 2, in which images with opposite perceptions are passed into the model. The naive cosine similarity for single image text pair is adjusted to

$$s_i = \frac{\mathbf{x} \odot \mathbf{t}_i}{\|\mathbf{x}\| \cdot \|\mathbf{t}_i\|}, \quad i \in \{1, 2\}, \tag{4}$$

where t_1 and t_2 are categories with opposing meanings (e.g., complex/simple, natural/synthetic, happy/sad, scary/peaceful, and new/old). This tweak effectively helped to enhance the model's conceptual understanding of abstract concept.

To make the model resolution-invariant, Wang et al. [1] removed the commonly used positional embedding in CLIP, which indicates the spatial relations between image patches in the image encoder and sequential relations between texts in the text encoder. The authors indicated that for perception assessment, such non-conceptual information will not help the training much. The authors also conducted comparative analysis on the backbones of CLIP. They found that ResNet is less sensitive to the removal of positional embeddings than transformers, and ended up choosing a ResNet-50 without positional embeddings.

4.3 Evaluation

Wang et al. [1] conducted a user study to test how well CLIP-IQA works for evaluating abstract perceptions. In the study, the authors created 15 pairs of images for each of five abstract perception attributes (i.e., complex/simple, natural/synthetic, happy/sad, scary/peaceful, and new/old). The image with a higher score was labeled as "positive" (e.g., happy), while the one with a lower score was labeled as "negative" (e.g., sad). A total of 25 participants rated each pair by selecting the image that best matched a provided description (e.g., "Which image makes you feel happier?"). The authors then compared CLIP-IQA's predictions to the participants'. CLIP-IQA achieved about 80% accuracy across all five attributes. It is worth noting that the authors did not publish specific rating criterion for each abstract attribute though (e.g., what is counted as old). The authors did not make their human-annotated test dataset publicly available, which led us to curate our own dataset, which is detailed in the next section.

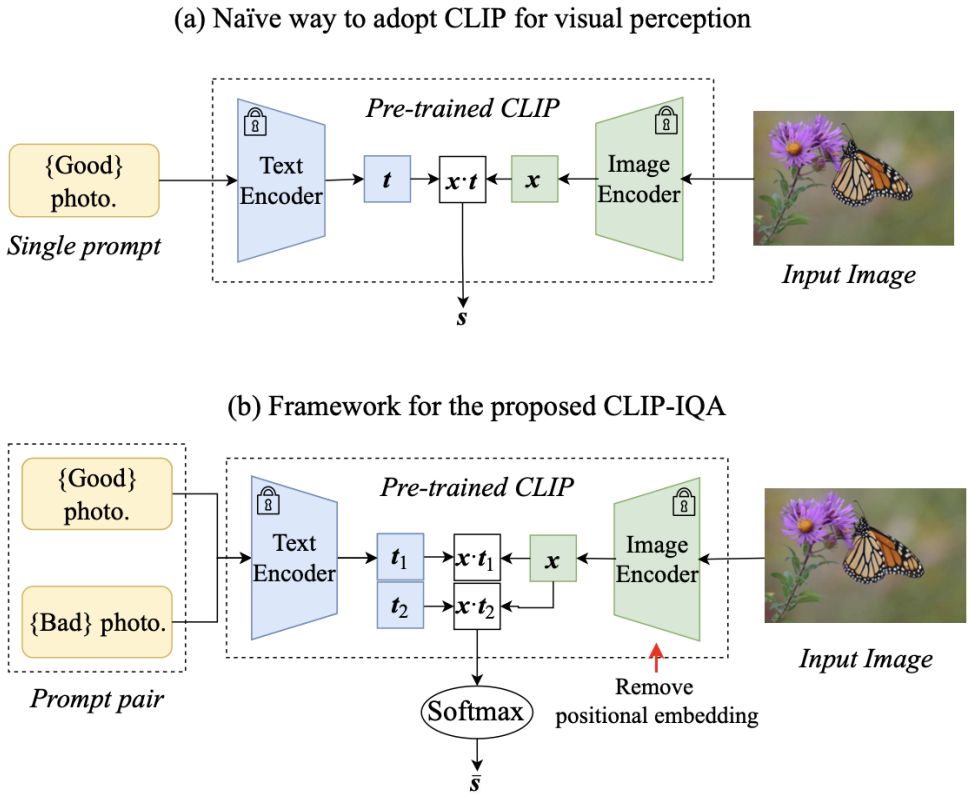

Figure 2: A naive approach of using a single prompt for perception assessment. (b) CLIP-IQA approach with (1) antonym prompt pairing strategy and (2) positional embedding removed.

5 Baseline Implementation

5.1 Dataset

5.1.1 AVA

The dataset used for our baseline implementation is the Aesthetic Visual Analysis (AVA) dataset [5], introduced to address the challenges of aesthetic evaluation in computer vision. With over 250,000 images sourced from an online photography community, AVA provides aesthetic scores generated by community votes, as well as semantic and photographic style labels, capturing both objective and subjective visual qualities. AVA's extensive metadata—spanning 60+ categories and 14 unique photographic styles—enables models to train on complex subjective attributes, making it suitable for evaluating intangible aspects like aesthetics and emotion. Notably, CLIP-IQA [1] leverages AVA to assess abstract qualities, validating its versatility for abstraction perception tasks.

5.1.2 KonIQ-10k

The KonIQ-10k dataset is used to train our CLIP-UIQA model. KonIQ-10k is a benchmark designed to evaluate the performance of models like CLIP-IQA, specifically within the domain of no-reference image quality assessment (NR-IQA). Featuring a diverse collection of images that capture realistic camera distortions, KonIQ-10k provides an ecologically valid framework for assessing image quality without requiring reference images. This dataset plays a pivotal role in benchmarking the effectiveness of various image quality evaluation methods, encompassing both traditional non-learning-based techniques and advanced learning-based approaches, including CLIP-IQA.

5.2 Evaluation

Figure 3: lawn and buildings in the frame of an arc

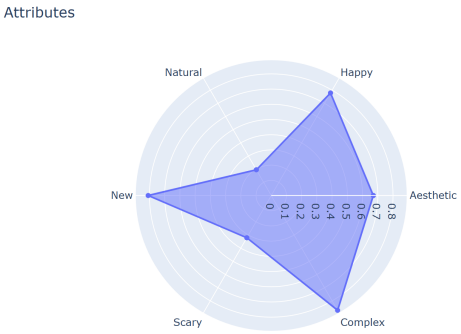

Figure 4: perception rating

Figure 5: a spooky doll in classic dress sitting on a sofa

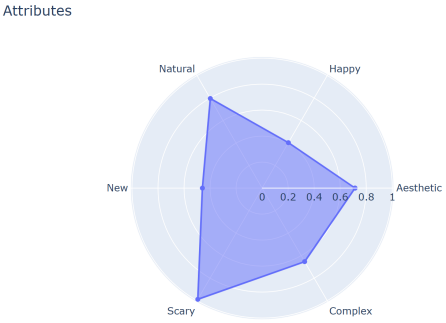

Figure 6: perception rating

Figure 7: flowers and leaves against the sky from a bottom-up shooting angle

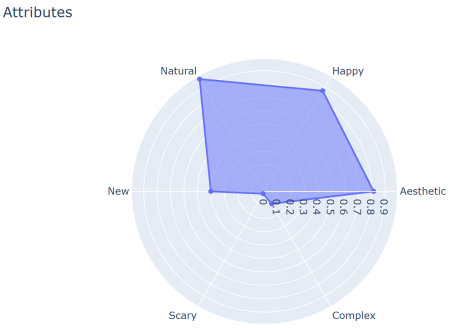

Figure 8: perception rating

5.2.1 Abstract Perception

To evaluate the model, we created a customizable script to perform abstract perception assessment with CLIP-IQA on a single text image. More specifically, the script allows the user to input a set of abstract attributes they want to evaluate and the backbone model they want to use. The script processes a single image, evaluates it on multiple aesthetic and perceptual attributes, and generates a radar (polar) plot showing the model's ratings for each attribute. Please see Figure 5.2 for some vivid examples in detail.

We successfully reproduced the model weights by deploying its most recent checkpoint and debugging some of the system glitches. Since the authors of CLIP-IQA did not make their abstract perception assessment user test dataset available, we decided to create our own dataset. We used the model to predict 100 randomly selected images. Four people in this team manually labeled each of six attributes (complex/simple, natural/synthetic, happy/sad, scary/peaceful, aesthetic/ugly, and new/old)

on a scale of 0.0 to 1.0. The higher the score is, the annotator agrees with the former in a pair of attributes listed above more.

We then performed Pearson's Linear Correlation Coefficient (PLCC) analysis on each of the six categories, as it is one of the analysis adopted in [1].

5.2.2 Image Quality

Image quality assessment focuses on evaluating objective aspects such as sharpness, noise, and structural integrity. For this, we utilized the KonIQ-10k dataset, which contains over 10,000 real-world images annotated with Mean Opinion Scores (MOS), providing a reliable benchmark for technical quality evaluation.

Our evaluation used standard metrics, including PSNR (distortion), SSIM (structural similarity), and PLCC/SRCC (correlation with MOS). Baseline results from CLIP-IQA demonstrated competitive performance, but limitations in addressing diverse distortions highlighted areas for improvement.

Our proposed model, CLIP-UIQA, improves on these metrics by incorporating a multi-branch architecture that captures global, local, and salient features, delivering a more nuanced understanding of image quality. Experimental results showed faster convergence and better alignment with human scores, positioning our model as a robust solution for image quality assessment.

6 Implemented Extensions/Experiments

6.1 Fine-Tune

Since the abstract perception score we got differ significantly from what is reported in the paper (please see Results/Abstract Perception for explanation), also because CLIP is a huge model and we have limited computing resources, we focused on enhancing image quality for the implemented extensions. To enhance CLIP-IQA's performance and adaptability, we conducted intensive hyper-parameter search and trained for intensive number of iterations to reach a comparable performance. Different variations of the models are fine-tuned specifically for image quality assessment tasks. The fine-tuning process involved:

1. **Optimized Prompts**: Tailored prompts were designed to guide the model in evaluating both technical and subjective attributes, such as "Good photo" vs. "Bad photo" for quality, and "Happy" vs. "Sad" for mood.
2. **Dataset-Specific Training**: The model was fine-tuned on datasets like KonIQ-10k for technical quality and our manually curated dataset for abstract attributes, ensuring alignment with diverse human perceptions.
3. **Loss Functions**: We experimented with different loss functions, such as Charbonnier for handling outliers and MSE for general quality predictions, to refine the model's sensitivity to subtle distinctions.
4. **Multitask Learning**: Fine-tuning also incorporated multitask objectives, enabling the model to learn shared representations for attributes like aesthetic appeal, mood classification, and distortion recognition.

	Iterations	Loss Function	Optimizer Configurations			Scheduler Configurations			
			Optimizer	LR	Weight Decay	Scheduler	Periods	Restart Weights	Minimum LR
Baseline	10000	MSE	SGD	0.002	-	CosineRestart	300000	1	1.00E-07
Variation 1	10000	MSE	AdamW	0.0002	0.01	CosineRestart	8000	1	1.00E-06
Variation 2	10000	Charbonnier	SGD	0.002	-	CosineRestart	300000	1	1.00E-07
Variation 3	10000	Charbonnier	AdamW	0.0002	0.01	CosineRestart	8000	1	1.00E-06

Figure 9: Hyperparameter Search on Baseline

6.2 CLIP-UIQA

We proposed a new architecture, combining CLIP-IQA and UIQA, a multi-branch DDN. While traditional Image Quality Assessment (IQA) models focus primarily on objective metrics such as sharpness and brightness, CLIP-UIQA introduces a multi-branch architecture that integrates diverse perspectives to evaluate images more comprehensively. We modified the whole pipeline of data processing to make sure each component in the pipeline, including rescale, normalization, flip, collect, etc, is compatible with intaking three formats of data. This augmented, comprehensive data processing pipeline sets CLIP-UIQA different from CLIP-IQA.

For the model architecture, CLIP-UIQA leverages three distinct branches, each processing a specific type of image input: Original images featuring global aesthetic characteristics , fragmented images featuring local technical distortion, and salient patches featuring the main focus.

More specifically, the salient patch is generated by central crop of tensors. After several manual inspections, we found that this straightforward strategy effectively captures the key, semantically rich portions of randomly selected images, mainly because most photographers adopt a central or slightly off-center focus.

The fragmented image is created by dividing the original image into spatially aligned or randomly shifted fragments of specified dimensions. Based on the number of fragments required horizontally and vertically, it dynamically calculates the size of each fragment. It also incorporates an upsample feature with bilinear interpolation, upsampling images whose resolutions are too low to be segmented. The starting position of each fragment is offset by a random variable. This creates the effect that some fragments adhere to grid positions while others overlap or have a jittery effect.

Instead of the Swin Transformer backbone used in UIQA, for each of the three branches, the model uses a CLIP Predictor to generate high-quality embeddings. We also modified the CLIP-IQA image encoder to make it not only multi-branch for different classes (e.g., good/bad photo and happy/sad photo are two classes) but also multi-branch for different image formats. We also experimented with how the model should holistically evaluate and aggregate information from all the branches. We added a mechanism that the outputs from the three image format branches are aggregated—either through averaging or concatenation. The output is then concatenated with outputs from different classes, before being passed to a non-linear regression head that predicts the final quality score.

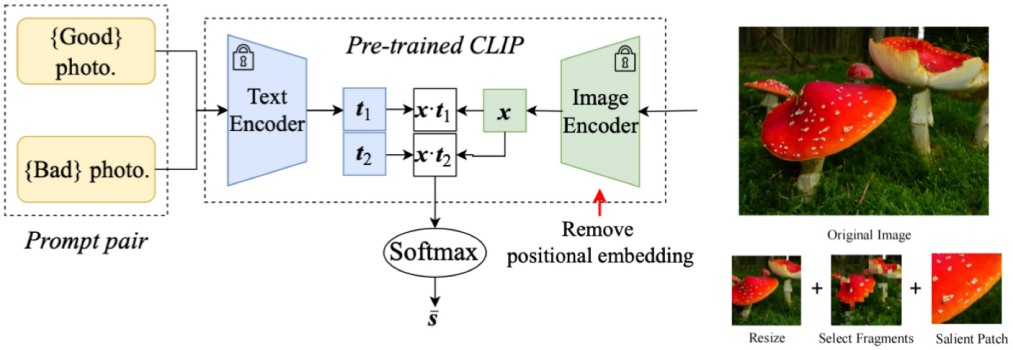

Figure 10: data and image preprocessing pipeline of CLIP-UIQA

6.3 Manually Curated Dataset

To address limitations in existing datasets, we created a manually curated dataset designed to capture subjective attributes like mood, emotion, and aesthetic appeal. This dataset contains approximately 700 images, ensuring a diverse set of styles, themes, and contexts. This dataset includes diverse images annotated with descriptors including "natural", "happy," "aesthetics", "new", "complex", "scary," and "image quality," focusing on both commonly agreed and context-dependent attributes. Each image was carefully selected and labeled through group reviews to ensure balanced representation and alignment with human perceptions. This curated dataset serves as a foundation for refining our model's ability to interpret subtle, subjective elements. The dataset can be found in our git repository link: `https://github.com/hailsong/CLIP-IQA`

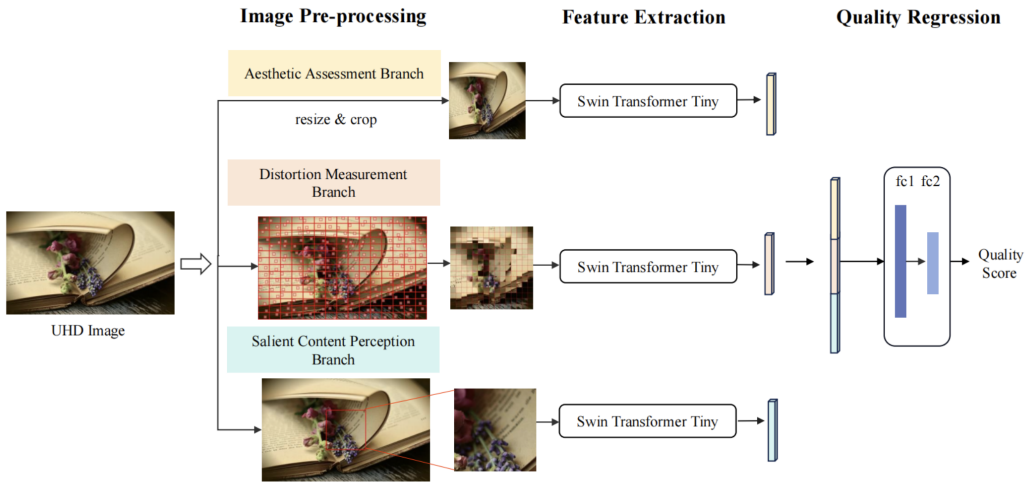

Figure 11: model architecture of CLIP-UIQA

7 Results

7.1 Abstract Perception

We used the model to predict abstract perceptions and conducted Pearson analysis with the ground truth. The results are listed in Table 1

Attribute Pair	Pearson Score
Complex/Simple	0.20
Natural/Synthetic	0.43
Happy/Sad	0.10
Scary/Peaceful	0.20
Aesthetic/Ugly	0.41
New/Old	0.08

Table 1: Attribute pairs and their Pearson scores

We can see the two comparably accurate prediction categories are Natural/Synthetic and Aesthetic/Ugly, slightly above 40%, while the others are around 10% - 20%. One of the explanations for such discrepancy is sample bias. While the original paper conducted a user study, they did not publish the set of test images they used so we curated and annotated a test dataset on our own. A hundred images each with six labels on an eleven-point likert scale (0.0 - 1.0), though intensive for us to manually label, is a really small test dataset for machine learning models. Abstract perceptions are also intrinsically subjective. Even the perceptions between two annotators of this project on the same image could be drastically different, not to mention a different group users on a different dataset. Since Wang et al. [1] did not publish their user evaluating criterion, it is hard to produce the ratings. We still believe our evaluation has some values. It sheds light on the problem that the performance of state-of-the-art abstract perception models is not as optimistic as reported in the literature, which highlights the importance of this research project. In addition, our manually labeled dataset can be used as new training points to further augment the model. Furthermore, the superior performance of categories Natural/Synthetic and Aesthetic/Ugly, compared with the rest of categories, leaves fruitful research questions for us to investigate: why does CLIP-IQA excel in those two categories? What is the model is learning in those two categories? For example, is the model actually learning the abstract concept or is it learning special objects in those categories (e.g., grass for nature)? Can we borrow insights from those two categories to enhance the performance of the rest? We will delineate our future steps in the next section.

7.2 Image Quality

CLIP-IQA demonstrated better performance than the baseline. Especially in early iterations. It quickly converges in the first 100 iterations to what the original CLIP-IQA doesn't reach until 400 iterations.

Table 2: Best Losses

Variant	Iterations	Loss Function	Optimizer	LR	Scheduler	Periods	Best Loss
Baseline	10,000	MSE	SGD	0.002	CosineRestart	300,000	0.0047
Baseline w/ Char loss	10,000	Charbonnier	SGD	0.002	CosineRestart	300,000	0.0481
CLIP-UIQA	10,000	MSE	SGD	0.002	CosineRestart	300,000	0.0045

8 Discussion

Our work demonstrates the potential of integrating vision-language models like CLIP with a multi-branch architecture to assess both the technical and abstract qualities of images. While the baseline model effectively captures certain subjective attributes, such as "happy" and "scary," challenges remain in interpreting context-dependent or vague attributes. These inconsistencies highlight the inherent difficulty of quantifying subjective human perceptions.

One of the reasons we thought CLIP-UIQA converged fast and worked well is because it combines the advantages of both CLIP-IQA and UIQA. Compared with CLIP-IQA, the augmented image processing pipeline emphasizes which part of image the model should focus on, similar to human attention, using the salient patch. In other words, the model doesn't need to spend hours learning in an image what the essential information is and what impacts human perceptions the most. The fragmented images also add much more variability to the dataset, leading to better transfer learning and generalization. Compared with UIQA, which uses Swin Transformer as its backbone, CLIP-UIQA leverages the powerful model CLIP. While Swin is a vision-only model, CLIP, on the other hand, is an vision-text multi-class model. Cross-model learning enables CLIP to understand visual content in the context of natural language, which is also what humans use to express subjective, abstract perceptions. The paired antonym prompts added in CLIP-IQA allow it to capture precisely what image quality or abstract perception one wants to value. Since CLIP is pretrained on massive datasets, it generalizes well to a wide range of tasks without task-specific fine-tuning, while Swin usually requires much more fine-tuning for downstream tasks.

The results also underline the importance of carefully crafted prompts and attribute definitions. Subtle variations in phrasing significantly influence how models interpret abstract qualities, emphasizing the need for a more robust prompting strategy. Additionally, the limitations in subjective evaluation underscore the influence of cultural and personal factors, suggesting that future models may benefit from adapting to user-specific contexts.

Looking ahead, this work opens opportunities for developing models that better align with human aesthetic judgments. As machine perception continues to evolve, integrating contextual, cultural, and emotional factors will be key to creating systems that truly understand and replicate human-like evaluations.

9 Future Works

Our future work aims to build upon the baseline to better encapsulate subjective aspects of human visual perception. This involves integrating optimized prompting strategies and implementing a multitask learning framework to enhance abstract perception capabilities in vision-language models like CLIP.

9.1 Tailored Prompts for Abstract Attributes

For future design, we suggest designing and evaluating tailored prompts that specifically prioritize the emotional and aesthetic dimensions of an image. For example, we propose developing context-sensitive prompts that emphasize subtle distinctions aligned with human perception and incorporating feedback loops to iteratively refine prompt construction based on model misclassifications.

These prompts will guide the model in evaluating intangible qualities such as mood, appeal, and aesthetic harmony. Current baseline results show promising performance for attributes like "happy" and "scary." However, challenges persist in capturing attributes such as "natural," where context-dependent synthetic elements, like a spooky doll, are misclassified. Subjective and vague characteristics, such as "new" and "complex," are inconsistently interpreted, highlighting the need for more refined attribute definitions and targeted prompting strategies.

9.2 Enhanced Multitask Learning Framework

We plan to extend the multitask learning framework to capture complex, context-dependent perceptions. By training the model on interrelated tasks, such as mood classification, aesthetic rating, and emotion detection, we aim to foster deeper, contextually informed interpretations of images. The multitasking framework will enable the model to learn shared representations that connect diverse abstract attributes and improve the model's ability to generalize across subjective dimensions by emphasizing contextual cues.

Currently, we equally weigh the output embeddings from the three image format branches. In the future, we want to explore weighted average or more dynamic weighting strategies that emphasize or de-emphasize different formats of images at different stages of the data pipeline. We also want to explore how we can more dynamically aggregate information from different branches, beyond concatenation and averaging. While CLIP-UIQA focused more on the image quality side, we want to do more testing with abstract perception with our new manually curated dataset.

9.3 Proposed Evaluation Strategies

To evaluate the effectiveness of these enhancements, we will benchmark the refined model against the baseline. Key evaluation strategies include: - Using datasets that incorporate diverse subjective ratings and challenging abstract attributes. - Comparing model predictions with human ratings using correlation metrics, such as Spearman's rank correlation, to assess alignment with subjective human judgments. - Testing the model's robustness on real-world applications, such as personalized content recommendation and automated media curation, to validate its practical utility.

We hypothesize that these extensions will significantly improve the model's performance in subjective assessments and bring its outputs close to human judgments, broadening its utility in real-world applications such as content recommendation, automated curation, and user-centric aesthetic assessments. Evaluation of our proposed extensions will involve benchmarking against the baseline model with an emphasis on comparing the models' predictions to human ratings using correlation metrics, such as Spearman's rank correlation, to assess alignment with subjective human judgments.

10 Conclusion

In this work, we introduced CLIP-UIQA, a multi-branch model that integrates vision-language capabilities with enhanced architectural design to assess both technical and subjective image qualities. By combining global, local, and salient perspectives, our approach effectively bridges the gap between machine assessment and human perception. Fine-tuning with optimized prompts and multitasking learning further refined the model's ability to capture abstract attributes like mood and aesthetics.

Experimental results demonstrated that CLIP-UIQA outperforms baseline models in both convergence speed and alignment with human judgments while identifying areas for improvement in subjective evaluation. Future work will focus on refining prompt strategies, addressing cultural and contextual biases, and expanding datasets to enhance the model's adaptability and real-world applications.

Our contributions lay the groundwork for advancing machine perception, offering promising tools for content recommendation, automated curation, and user-centric aesthetic evaluations.

11 Division of Work

Eryn is responsible for labeling 400 data entries, writing up the math description and evaluation metrics of the baseline description, and writing up the evaluation of the baseline implementation (just the write-up not the implementation itself). Eryn developed the CLIP-UIQA framework, conducted

initial testings on it, and wrote up sections involving CLIP-UIQA (implemented extension, discussion, and future work).

Yuchen is responsible for drafting and revising the introduction and literature review, adding and revising key sections including future work, discussion, and CLIP-UIQA, adding a handcrafted dataset of 500 labeled images, detailing fine-tuning methods, and drafting a comprehensive conclusion to enhance clarity, depth, and practical impact.

Shiwon is responsible for conducting the experiments and producing the results, drafting and revising the introduction, literature review, baseline selection, dataset, and proposed model extension, and also contributed to labeling the test data.

Hail is responsible for the coding work related to the baseline implementation and evaluation, including the visualization of Figures 3 to 8, and also contributed to labeling the test data.

References

[1] 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*, volume 37, pages 2555–2563, 2023.

[2] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14071–14081, 2023.

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

[4] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. 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*, pages 3667–3676, 2020.

[5] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2408–2415. IEEE, 2012.

[6] Simon Hentschel, Konstantin Kobs, and Andreas Hotho. Clip knows image aesthetics. *Frontiers in Artificial Intelligence*, 5:976235, 2022.

[7] Liwu Xu, Jinjin Xu, Yuzhe Yang, Yijie Huang, Yanchun Xie, and Yaqian Li. Clip brings better features to visual aesthetics learners. *arXiv preprint arXiv:2307.15640*, 2023.

[8] Hanna-Sophia Widhoelzl and Ece Takmaz. Decoding emotions in abstract art: Cognitive plausibility of clip in recognizing color-emotion associations. *arXiv preprint arXiv:2405.06319*, 2024.

[9] Guolong Wang, Yike Tan, Hangyu Lin, and Chuchun Zhang. Keep knowledge in perception: Zero-shot image aesthetic assessment. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8311–8315. IEEE, 2024.