The Photographer's Eye: Teaching Multimodal Large Language Models to See and Critique like Photographers (Appendix)

Anonymous CVPR submission

Paper ID 14571

A. PhotoCritique

In this section, we introduce details of PhotoCritique.

A.1 Construction of PhotoCritique

The PhotoCritique dataset contains 2.6M instructing-tuning samples. Each sample consists of the following elements:

- A URL to an image.
- A question.
- An answer to the question.

Based on their type, these instruction-tuning samples are categorized into three groups: aesthetic description, aesthetic conversation, and aesthetic visual question answering (VQA). We introduce each of them in the following.

The high-level data generation pipeline involves the following steps:

1. **Data Acquisition:** We obtain the raw data from DPC2022 [4], a publicly available collection of images and their corresponding raw comments from DPChallenge.
2. **Instruction-tuning Sample Generation:** For each image, question-answering samples are generated based on user comments using large language models (LLMs).

A.1.1 Generation of Aesthetic Description.

The generation of aesthetic description consists of two steps: critique generation and filtering. We prompt LLMs to perform each of the tasks. We summarize input, output and the involved LLM during each of the step in the following.

Step.1: Summarize a single paragraph of photo critique given a list of raw comments of an image.

- **Input:** A list of raw comments about an image and a predefined aesthetics-related question.
- **LLM:** LLaMA-3-70B-Instruct.
- **Output:** A single summarized comprehensive photo critique of the input image.

An example of this process is illustrated in Fig. 1.

Step 2: Filter the generated critiques from the previous step.

- **Input:** A single photo critique of an image.
- **LLM:** LLaMA-3-8B-Instruct.
- **Output:** A qualification indicator specifying whether the critique meets the quality requirements.

We only accept critique that meet the quality requirements, and reject unqualified ones. The aesthetic description part of PhotoCritique consists of all accepted samples from the process.

A.1.2 Generation of Aesthetic Conversation.

The generation of aesthetic conversation follows directly from the previous step. All accepted critiques from the earlier process are utilized to create aesthetic conversations.

The generation of aesthetic conversation consists of two steps: conversation generation and filtering. We prompt LLMs to perform each of the tasks. We summarize input, output and the involved LLM during each of the step in the following.

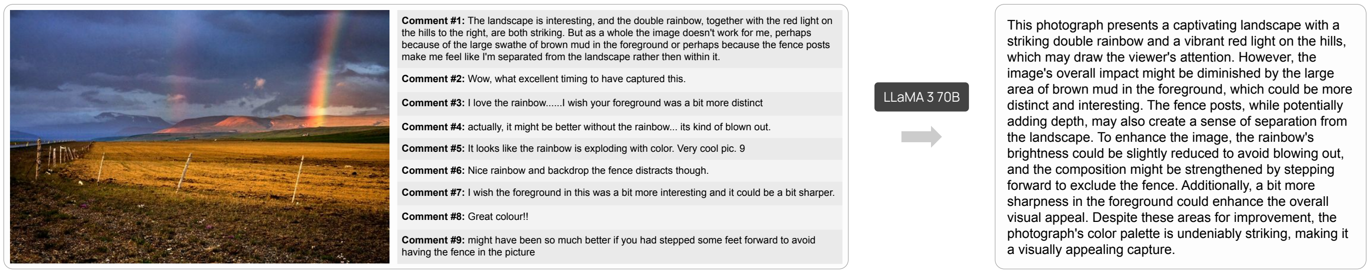
Step.1: Given a sampled aesthetics-related question and a photo critique, LLM is asked to generate response to the question based on the ground truth critique.

- **Input:** A sampled aesthetics-related question and a photo critique.
- **LLM:** LLaMA-3-70B-Instruct.
- **Output:** A response to the question according to the photo critique.

An example of this process is illustrated in Fig. 2.

Step 2: Filter the generated responses from the previous step. LLM is asked to tell whether the generated response is informative and correctly reflected in the critique.

- **Input:** A question, response, and the photo critique.
- **LLM:** LLaMA-3-8B-Instruct.
- **Output:** A qualification indicator specifying whether the response meets the quality requirements.



Comment #1: The landscape is interesting, and the double rainbow, together with the red light on the hills to the right, are both striking. But as a whole the image doesn't work for me, perhaps because of the large swathe of brown mud in the foreground or perhaps because the fence posts make me feel like I'm separated from the landscape rather then within it.

Comment #2: Wow, what excellent timing to have captured this.

Comment #3: I love the rainbow......I wish your foreground was a bit more distinct

Comment #4: actually, it might be better without the rainbow... its kind of blown out.

Comment #5: It looks like the rainbow is exploding with color. Very cool pic. 9

Comment #6: Nice rainbow and backdrop the fence distracts though.

Comment #7: I wish the foreground in this was a bit more interesting and it could be a bit sharper.

Comment #8: Great colour!!

Comment #9: might have been so much better if you had stepped some feet forward to avoid having the fence in the picture

This photograph presents a captivating landscape with a striking double rainbow and a vibrant red light on the hills, which may draw the viewer's attention. However, the image's overall impact might be diminished by the large area of brown mud in the foreground, which could be more distinct and interesting. The fence posts, while potentially adding depth, may also create a sense of separation from the landscape. To enhance the image, the rainbow's brightness could be slightly reduced to avoid blowing out, and the composition might be strengthened by stepping forward to exclude the fence. Additionally, a bit more sharpness in the foreground could enhance the overall visual appeal. Despite these areas for improvement, the photograph's color palette is undeniably striking, making it a visually appealing capture.

Raw Comments from DPChallenge (9 samples are shown)

Critique

Figure 1. An example of the generation of aesthetic description (step one). Note that **only raw comments** are input to the LLM. Image is only for illustration purpose.

We only accept responses that meet the quality requirements, and reject unqualified ones. The aesthetic conversation part of PhotoCritique consists of all accepted samples from the process.

A.1.3 Generation of Aesthetic VQA.

The generation of aesthetic VQA also follows directly from the Section A.1.1. All accepted critiques from the earlier process are utilized to create aesthetic VQA.

The generation of aesthetic VQA consists of two steps: multiple choice questions (MCQ) generation and filtering. We prompt LLMs to perform each of the tasks. We summarize input, output and the involved LLM during each of the step in the following.

Step.1: Given a photo critique, LLM is asked to generate 5 MCQ based on the ground truth critique.

- **Input:** A photo critique.
- **LLM:** LLaMA-3-70B-Instruct.
- **Output:** Five multiple choice questions with the ground truth answer labeled.

Step 2: Filter the generated MCQ from the previous step. LLM is asked to tell whether the generated MCQ and the ground truth answer is informative and correctly reflected in the critique.

- **Input:** MCQ and the photo critique.
- **LLM:** LLaMA-3-8B-Instruct.
- **Output:** A qualification indicator specifying whether the MCQ meets the quality requirements.

We only accept the MCQ that meets the quality requirements, and reject unqualified ones. The aesthetic VQA part of PhotoCritique consists of all accepted MCQ from the process.

A.2 More Examples of PhotoCritique

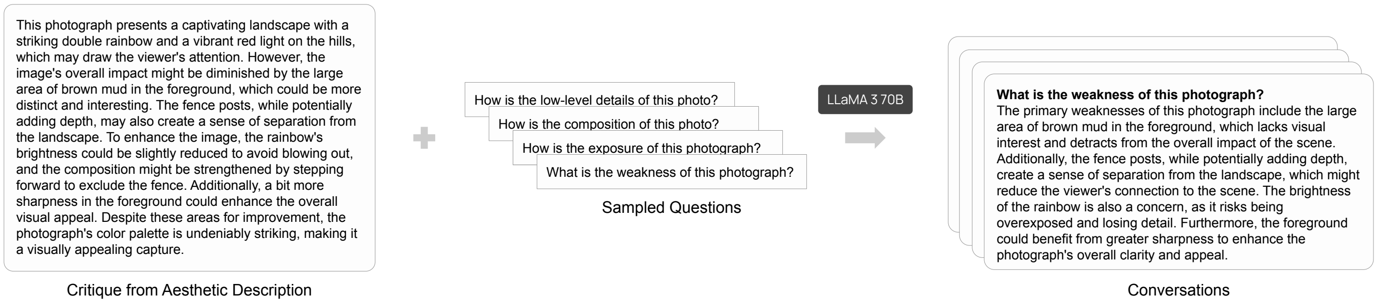
We presented more examples in Fig. 4. The dataset samples demonstrate high quality through their detailed content and analysis: **Rich and Context-Specific Descriptions**: Each critique offers tailored insights, addressing specific features of the images. For example, the *S-curve* in the first photo is praised for guiding the viewer's eye, while the *sepia tone* enhances the *nostalgic and timeless atmosphere*. This level of detail demonstrates a deep understanding of aesthetic principles. **Balanced Evaluation of Strengths and Weaknesses**: The critiques highlight both strengths and potential distractions, ensuring a nuanced analysis. For instance, the parasailing photo is commended for its *sense of drama and energy* created by the *warm glow*, while noting the *dark corner at the bottom left* might distract viewers. **Integration of Subjective and Objective Elements**: The critiques seamlessly combine subjective impressions, such as mood and narrative, with objective evaluations of composition and technique. For example, the urban photo's *dramatic atmosphere* is linked to its *high-contrast black-and-white treatment* and *skillful framing*. **Focus on Emotional and Artistic Impact**: Emotional resonance is a key feature of these critiques, such as the *lightness and freedom* conveyed by the parasailing photo or the *moody and introspective* tone of the urban scene. These observations demonstrate the dataset's emphasis on capturing artistic intent. **Clarity and Professional Tone**: The critiques are generated with precision and professionalism, avoiding vague language. Our dataset excels in delivering detailed critiques that combine technical evaluation with artistic interpretation, making it a valuable resource for research in image aesthetics.

B. Learning Aesthetic Visual Features with PhotoEye

In this section, we conduct a deeper investigation into the learning of aesthetic concepts at feature level with the vision module in PhotoEye, which lays the foundation of high-quality aesthetic response from LLM.

B.1 Discriminability of Aesthetic Visual Features

Most existing works [1, 3] use CLIP as the vision encoder, which, however, is pre-trained with high-level image-text alignment in general domains and is relatively less effective at capturing aesthetic elements. Fig. 1 (right) in the


This photograph presents a captivating landscape with a striking double rainbow and a vibrant red light on the hills, which may draw the viewer's attention. However, the image's overall impact might be diminished by the large area of brown mud in the foreground, which could be more distinct and interesting. The fence posts, while potentially adding depth, may also create a sense of separation from the landscape. To enhance the image, the rainbow's brightness could be slightly reduced to avoid blowing out, and the composition might be strengthened by stepping forward to exclude the fence. Additionally, a bit more sharpness in the foreground could enhance the overall visual appeal. Despite these areas for improvement, the photograph's color palette is undeniably striking, making it a visually appealing capture.

How is the low-level details of this photo?

How is the composition of this photo?

How is the exposure of this photograph?

What is the weakness of this photograph?

What is the weakness of this photograph?
The primary weaknesses of this photograph include the large area of brown mud in the foreground, which lacks visual interest and detracts from the overall impact of the scene. Additionally, the fence posts, while potentially adding depth, create a sense of separation from the landscape, which might reduce the viewer's connection to the scene. The brightness of the rainbow is also a concern, as it risks being overexposed and losing detail. Furthermore, the foreground could benefit from greater sharpness to enhance the photograph's overall clarity and appeal.

Figure 2. An example of the generation of aesthetic conversation (step one). Note that **only critiques** are input to the LLM. Image is only for illustration purpose.

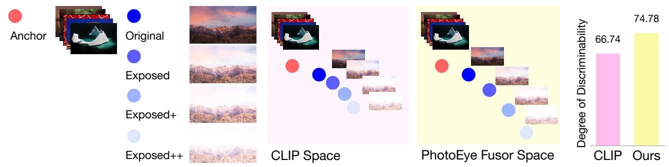


Figure 3. Comparison of aesthetics-related visual feature discriminability of ours and CLIP [2] in existing works [1, 3, 5].

body of the paper (main text) illustrates this issue, where Q-Instruct [3] and AesExpert [1] only report 'overexposure' when the image is severely overexposed.

Compared to existing works, PhotoEye incorporates a more powerful multi-view vision fusor to extract more fine-grained aesthetics-related visual features. We designed an experiment to explore the discriminability of visual features extracted by our vision module and those from existing works from an aesthetic perspective.

Specifically, for a given image, we first select a low-level aesthetic attribute and edit the original image to generate a series of edited images. For example, as shown in Fig. 3, we gradually increase the exposure of the original image to create a sequence of edited images.

With the original image and the edited images, we compute the average distance between the original image and edited images in the feature space. Intuitively, greater distance between the original image and edited images in feature space indicates that the vision encoder is more sensitive to this low-level aesthetic attribute. In other words, it indicates changes in this attribute in the pixel space lead to more significant changes in the feature space, making them more discriminative, which is desirable for aesthetic visual understanding.

However, the challenge is that, the absolute value of the average distance between the original and edited images cannot be directly compared across different feature spaces. This is due to differences in the overall feature density of each feature space. To address this issue, we propose comparing the *relative distance*.

Specifically, for a given original image, we randomly sample a certain number of anchor images from PhotoCritique (e.g., 1,000) and calculate the average distance between the original image and these anchor images in the feature space. This serves as an approximate measure of the feature space's density. We then normalize the average distance between the original image and the edited images with this value. Denoting the feature of the original image as $\mathbf{X}_\mathrm{o}$, the edited images as $\{\mathbf{X}_\mathrm{e}^i\}_{i=1}^N$ (assuming there are N edited images), and the anchor images as $\{\mathbf{X}_\mathrm{a}^i\}_{i=1}^M$ (assuming there are M anchor images), we define the *relative distance* of the original image and edited images as the degree of discriminability:

$$\textit{Relative Distance} = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathrm{DIST}(\mathbf{X}_o,\mathbf{X}_e^i)}{\frac{1}{M}\sum_{j=1}^{M}\mathrm{DIST}(\mathbf{X}_o,\mathbf{X}_a^j)}, \tag{1}$$

where DIST() is the distance function, and we use cosine distance (measured in degrees) in this case.

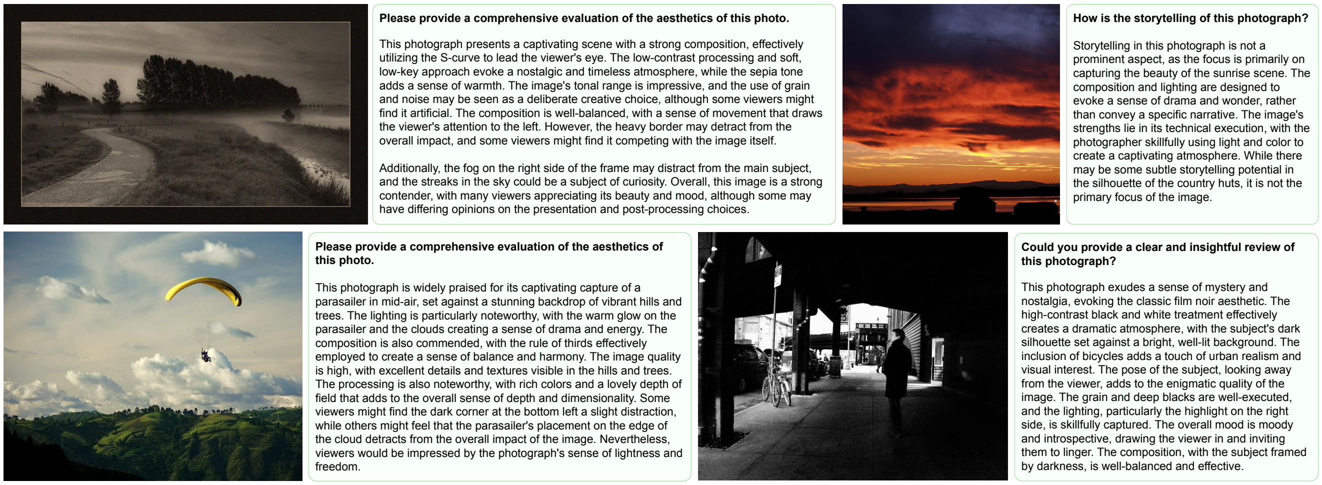
In this way, we can compare the *relative distances* between the original image and the edited images across different feature spaces. A larger value indicates that the vision encoder is more sensitive to changes in the specific low-level aesthetic attribute of the original image, indicating the vision encoder is more sensitive to this aesthetic attribute.

In Fig. 3, we show the example in terms of exposure. In experiments, we use $M = 100$ and $N = 3$, and randomly sampled 50 images as the original images from PhotoCritique. Results in Fig. 3 (right) show that CLIP is less effective in discriminating towards aesthetics-related features as they are more densely clustered, making them less discriminative. By contrast, the vision module of PhotoEye is more sensitive to it, yielding more discriminative features.

C. PhotoBench

The PhotoBench consists of multiple choice questions. In each sample, there are:

- A URL to an image.



Please provide a comprehensive evaluation of the aesthetics of this photo.

This photograph presents a captivating scene with a strong composition, effectively utilizing the S-curve to lead the viewer's eye. The low-contrast processing and soft, low-key approach evoke a nostalgic and timeless atmosphere, while the sepia tone adds a sense of warmth. The image's tonal range is impressive, and the use of grain and noise may be seen as a deliberate creative choice, although some viewers might find it artificial. The composition is well-balanced, with a sense of movement that draws the viewer's attention to the left. However, the heavy border may detract from the overall impact, and some viewers might find it competing with the image itself.

Additionally, the fog on the right side of the frame may distract from the main subject, and the streaks in the sky could be a subject of curiosity. Overall, this image is a strong contender, with many viewers appreciating its beauty and mood, although some may have differing opinions on the presentation and post-processing choices.

How is the storytelling of this photograph?

Storytelling in this photograph is not a prominent aspect, as the focus is primarily on capturing the beauty of the sunrise scene. The composition and lighting are designed to evoke a sense of drama and wonder, rather than convey a specific narrative. The image's strengths lie in its technical execution, with the photographer skillfully using light and color to create a captivating atmosphere. While there may be some subtle storytelling potential in the silhouette of the country huts, it is not the primary focus of the image.

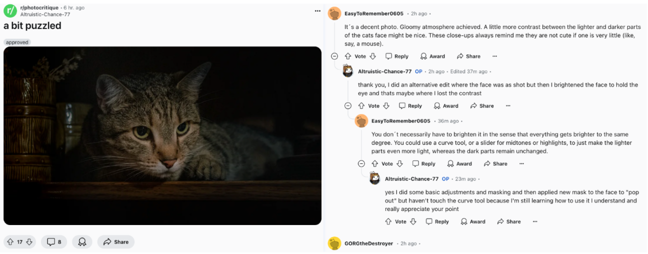
Please provide a comprehensive evaluation of the aesthetics of this photo.

This photograph is widely praised for its captivating capture of a parasailer in mid-air, set against a stunning backdrop of vibrant hills and trees. The lighting is particularly noteworthy, with the warm glow on the parasailer and the clouds creating a sense of drama and energy. The composition is also commended, with the rule of thirds effectively employed to create a sense of balance and harmony. The image quality is high, with excellent details and textures visible in the hills and trees. The processing is also noteworthy, with rich colors and a lovely depth of field that adds to the overall sense of depth and dimensionality. Some viewers might find the dark corner at the bottom left a slight distraction, while others might feel that the parasailer's placement on the edge of the cloud detracts from the overall impact of the image. Nevertheless, viewers would be impressed by the photograph's sense of lightness and freedom.

Could you provide a clear and insightful review of this photograph?

This photograph exudes a sense of mystery and nostalgia, evoking the classic film noir aesthetic. The high-contrast black and white treatment effectively creates a dramatic atmosphere, with the subject's dark silhouette set against a bright, well-lit background. The inclusion of bicycles adds a touch of urban realism and visual interest. The pose of the subject, looking away from the viewer, adds to the enigmatic quality of the image. The grain and deep blacks are well-executed, and the lighting, particularly the highlight on the right side, is skillfully captured. The overall mood is moody and introspective, drawing the viewer in and inviting them to linger. The composition, with the subject framed by darkness, is well-balanced and effective.

Figure 4. Examples from PhotoCritique.



r/photocritique · 6 hr. ago
Altruistic-Chance-77

a bit puzzled

EasyToRemember0605 · 2h ago ·

It´s a decent photo. Gloomy atmosphere achieved. A little more contrast between the lighter and darker parts of the cats face might be nice. These close-ups always remind me they are not cute if one is very little (like, say, a mouse).

Altruistic-Chance-77 OP · 2h ago · Edited 37m ago ·

thank you, I did an alternative edit where the face was as shot but then I brightened the face to hold the eye and thats maybe where I lost the contrast

EasyToRemember0605 · 36m ago ·

You don´t necessarily have to brighten it in the sense that everything gets brighter to the same degree. You could use a curve tool, or a slider for midtones or highlights, to just make the lighter parts even more light, whereas the dark parts remain unchanged.

Altruistic-Chance-77 OP · 23m ago ·

yes I did some basic adjustments and masking and then applied new mask to the face to "pop out" but haven't touch the curve tool because I'm still learning how to use it I understand and really appreciate your point

GORGtheDestroyer · 2h ago ·

Figure 5. An example from Reddit Photocritique.

Hyperparameters	
Batch Size	128
Learning Rate	2×10^{-5}
Learning Rate Schedule	Cosine Decay
Learning Rate Warmup Ratio	0.03
Weight Decay	0
Epoch	1
Optimizer	AdamW
DeepSpeed Stage	3

Table 1. Hyperparameters for Instruction Finetuning.

- A question.
- Four answer choices, with one marked as the ground truth.
- A topic label describing the subject of the question.

We obtain the raw data from the Reddit Photo Critique Dataset (RPCD). RPCD is a collection of images and the corresponding raw photo comments from reddit, specifically r/photocritique. One example is available at Fig. 5. For each image, we generate 5 multiple-choice questions from only user comments with large language models. All generated multiple-choice questions are filtered by LLM.

The MCQ generation consists of the following steps:

Step 1: Summarize a single paragraph of photo critique given a list of raw comments of an image.

- **Input:** A list of raw comments of an image.
- **Model:** LLaMA-3-70B.
- **Output:** A single photo critique of an image.

Step 2: Generate five MCQ from a single photo critique of an image.

- **Input:** A single photo critique of an image.
- **Model:** GPT-4-turbo.
- **Output:** Five MCQ questions, and one label that describes the topic of the question.

*Step 3: LLM filtering: reject easy samples that the LLM can return the correct answer **without** referring to the image.*

- **Input:** Generated MCQ questions (text only).
- **Model:** GPT-4o.
- **Output:** The predicted answer.

In this step, we reject samples if the predicted answer is the ground truth answer.

Step 3: LLM filtering: rate generated questions based on aesthetics relevance, visual dependency, and expertise.

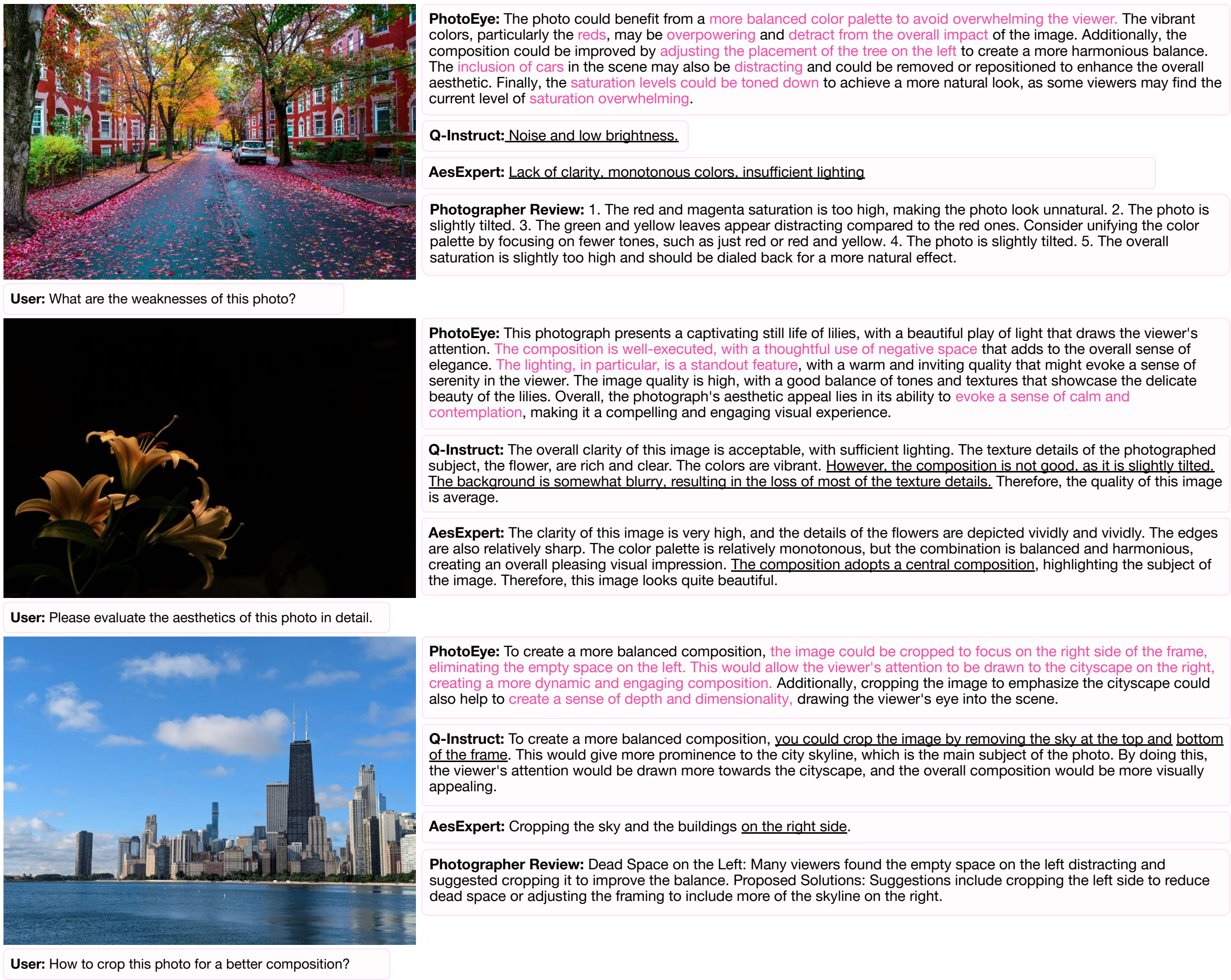
- **Input:** Generated MCQ questions (text only).
- **Model:** GPT-4o.
- **Output:** A score.

In this step, we selected the top 1,500 questions based on the average score.

D. Experiments

D.1. Hyperparameters for Instruction Finetuning

Details of hyperparameters for instruction finetuning are summerized in Tab. 1. The training takes around 96 hours



PhotoEye: The photo could benefit from a more balanced color palette to avoid overwhelming the viewer. The vibrant colors, particularly the reds, may be overpowering and detract from the overall impact of the image. Additionally, the composition could be improved by adjusting the placement of the tree on the left to create a more harmonious balance. The inclusion of cars in the scene may also be distracting and could be removed or repositioned to enhance the overall aesthetic. Finally, the saturation levels could be toned down to achieve a more natural look, as some viewers may find the current level of saturation overwhelming.

Q-Instruct: Noise and low brightness.

AesExpert: Lack of clarity, monotonous colors, insufficient lighting

Photographer Review: 1. The red and magenta saturation is too high, making the photo look unnatural. 2. The photo is slightly tilted. 3. The green and yellow leaves appear distracting compared to the red ones. Consider unifying the color palette by focusing on fewer tones, such as just red or red and yellow. 4. The photo is slightly tilted. 5. The overall saturation is slightly too high and should be dialed back for a more natural effect.

User: What are the weaknesses of this photo?

PhotoEye: This photograph presents a captivating still life of lilies, with a beautiful play of light that draws the viewer's attention. The composition is well-executed, with a thoughtful use of negative space that adds to the overall sense of elegance. The lighting, in particular, is a standout feature, with a warm and inviting quality that might evoke a sense of serenity in the viewer. The image quality is high, with a good balance of tones and textures that showcase the delicate beauty of the lilies. Overall, the photograph's aesthetic appeal lies in its ability to evoke a sense of calm and contemplation, making it a compelling and engaging visual experience.

Q-Instruct: The overall clarity of this image is acceptable, with sufficient lighting. The texture details of the photographed subject, the flower, are rich and clear. The colors are vibrant. However, the composition is not good, as it is slightly tilted. The background is somewhat blurry, resulting in the loss of most of the texture details. Therefore, the quality of this image is average.

AesExpert: The clarity of this image is very high, and the details of the flowers are depicted vividly and vividly. The edges are also relatively sharp. The color palette is relatively monotonous, but the combination is balanced and harmonious, creating an overall pleasing visual impression. The composition adopts a central composition, highlighting the subject of the image. Therefore, this image looks quite beautiful.

User: Please evaluate the aesthetics of this photo in detail.

PhotoEye: To create a more balanced composition, the image could be cropped to focus on the right side of the frame, eliminating the empty space on the left. This would allow the viewer's attention to be drawn to the cityscape on the right, creating a more dynamic and engaging composition. Additionally, cropping the image to emphasize the cityscape could also help to create a sense of depth and dimensionality, drawing the viewer's eye into the scene.

Q-Instruct: To create a more balanced composition, you could crop the image by removing the sky at the top and bottom of the frame. This would give more prominence to the city skyline, which is the main subject of the photo. By doing this, the viewer's attention would be drawn more towards the cityscape, and the overall composition would be more visually appealing.

AesExpert: Cropping the sky and the buildings on the right side.

Photographer Review: Dead Space on the Left: Many viewers found the empty space on the left distracting and suggested cropping it to improve the balance. Proposed Solutions: Suggestions include cropping the left side to reduce dead space or adjusting the framing to include more of the skyline on the right.

User: How to crop this photo for a better composition?

Figure 6. Real-world aesthetic conversation examples with MLLMs. Images are sampled from discussions on Reddit Photocritique. High-quality aesthetics-related descriptions are highlighted and vague or inaccurate descriptions are underlined.

with 8 A100 (80GB) GPUs with ZeRO2. We choose ZeRO2 to avoid some potential implementation and running issues with ZeRO3 in practice.

E. Qualitative Results

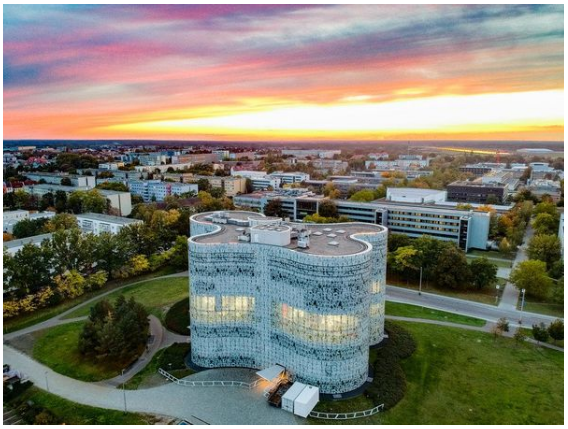
In this section, we present qualitative results of PhotoEye in different real-world scenarios and compare it with other models. Fig. 6 presents real-world examples, where images are sampled from discussions on Reddit Photocritique and we ask MLLMs to similar questions discussed on the platform.

PhotoEye demonstrates clear advantages in providing detailed, actionable feedback that include technical critique and aesthetic insight. For instance, it identifies specific issues like distracting elements or oversaturation while offering practical solutions, such as adjusting color balance or cropping to enhance composition. Compared to other models, it goes beyond surface-level observations, analyzing how artistic elements like lighting and space contribute to the emotional impact of a photo. This ability to provide nuanced, context-aware suggestions makes PhotoEye especially effective in real-world photographic critique. When discussing the weaknesses of the photos (as shown in the first and last conversations), the suggestions provided by PhotoEye closely align with those offered by professional photographers and enthusiasts from Reddit Photocritique.

F. Vision Encoders on Aesthetic Conversations

In the main text of the submission, we demonstrate that different vision encoders contribute differently under different scenarios. Specifically, we show that CoDETR plays a more important role in composition-related questions quantitatively.

In Fig. 7, we present an interesting real-world exam-



User: What are the weaknesses of this photo in detail?

PhotoEye: The main weakness of this photo is the overexposure of the sky, which may detract from the overall impact of the image. To improve this, the photographer could consider adjusting their exposure settings to balance the brightness of the sky with the rest of the scene, potentially using a graduated neutral density filter or adjusting the exposure compensation in post-processing. Additionally, the noticeable noise in the image could be reduced by using a lower ISO or applying noise reduction techniques in post-processing.

PhotoEye (Ablate CLIP): The composition of this photograph could be improved by adjusting the horizon to create a more balanced and visually appealing image. The current tilt of the horizon may be distracting to the viewer and detract from the overall impact of the photograph. Additionally, the lack of a clear subject or focal point may make it difficult for the viewer to engage with the image. To address this, the photographer could consider re-framing the shot to emphasize a specific element or adjusting the composition to draw the viewer's attention to a particular area of the image.

PhotoEye (Ablate CoDETR): The main weakness of this photo is the overexposure of the sky, which has resulted in a loss of detail in the brightest areas. To improve this, the photographer could consider adjusting their exposure settings to capture a more balanced range of tones, potentially using a graduated neutral density filter or adjusting the exposure compensation in post-processing. Additionally, the noticeable grain in the image could be reduced by using a lower ISO or applying noise reduction techniques in post-processing. Finally, the blown-out highlights in the sky could be addressed by adjusting the exposure or using a graduated neutral density filter to balance the contrast between the sky and the rest of the image.

AesExpert: The clarity of this image is very high, and the details of the buildings and the grass are clearly visible. The colors are very vibrant, with the orange-red sky and the green grass creating a sharp contrast. The lighting is very bright, with the sunlight shining on the buildings, creating a warm and soft atmosphere. The composition is very good, with the buildings and the grass occupying the foreground and the sky occupying the background, creating a sense of depth in the image. Therefore, this image looks very beautiful.

Q-Instruct: To further improve this image in detail, one could consider the following:

1. Enhance the clarity of the main subject, the building, by adjusting the focus and exposure.
2. Add more texture details to the main subject, such as the building's facade, by using sharpening or blurring techniques.
3. Improve the composition by adding more elements to the frame, such as trees, cars, or people, to create a more dynamic and interesting scene.
4. Adjust the lighting to create a more dramatic and visually appealing effect, such as using backlighting or overexposure to emphasize certain parts of the image.
5. Experiment with different color palettes and contrast levels to create a more vibrant and eye-catching image.

By applying these techniques, the quality of the image can be further enhanced, making it more visually appealing and engaging for viewers.

Photographer Review: This image may benefit from refinements in its HDR processing and noise reduction techniques. The viewer might find that the current stacked HDR approach, using only two images with limited spacing, results in an underwhelming effect, particularly in the blown-out highlight areas of the sky. Additionally, the image appears overly grainy, which detracts from the overall impact of the vibrant colors. To improve the image, it could be helpful to explore alternative HDR methods, such as using multiple shots with wider spacing, and employing noise reduction techniques in post-processing. Furthermore, the viewer might consider adjusting the shutter speed to capture more defined cloud details, rather than the current blurry streaks.

Figure 7. Real-world aesthetic conversation example with PhotoEye with different vision encoders activated. The image is sampled from discussions on Reddit Photocritique. High-quality aesthetics-related descriptions are highlighted and vague or inaccurate descriptions are underlined.

Dataset Type	In-the-wild				Generated		Artificial	
Model / Dataset	KONiQ-10k	SPAQ	LIVE-FB	LIVE-itw	CGIQA-6K	AGIQA-3K	KADID-10K	Average
NIQE	0.316/0.377	0.693/0.669	0.211/0.288	0.480/0.451	0.075/0.056	0.562/0.517	0.374/0.428	0.387/0.398
CLIP-ViT-Large-14	0.468/0.505	0.385/0.389	0.218/0.237	0.307/0.308	0.285/0.290	0.436/0.458	0.376/0.388	0.354/0.368
LLaVA-v1.5 (Vicuna-v1.5-7B)	0.463/0.459	0.443/0.467	0.305/0.321	0.344/0.358	0.321/0.333	0.672/0.738	0.417/0.440	0.424/0.445
LLaVA-v1.5 (Vicuna-v1.5-13B)	0.448/0.460	0.563/0.584	0.310/0.339	0.445/0.481	0.285/0.297	0.664/0.754	0.390/0.400	0.444/0.474
InternLM-XComposer-VL (InternLM)	0.564/0.615	0.730/0.750	0.360/0.416	0.612/0.676	0.243/0.265	0.732/0.775	0.546/0.572	0.541/0.581
IDEFICS-Instruct (LLaMA-7B)	0.375/0.400	0.474/0.484	0.235/0.240	0.409/0.428	0.244/0.227	0.562/0.622	0.370/0.373	0.381/0.396
Qwen-VL (QwenLM)	0.470/0.546	0.676/0.669	0.298/0.338	0.504/0.532	0.273/0.284	0.617/0.686	0.486/0.486	0.475/0.506
Shikra (Vicuna-7B)	0.314/0.307	0.320/0.337	0.237/0.241	0.322/0.336	0.198/0.201	0.640/0.661	0.324/0.332	0.336/0.345
Otter-v1 (MPT-7B)	0.406/0.406	0.436/0.441	0.143/0.142	-0.008/0.018	0.254/0.264	0.475/0.481	0.557/0.577	0.323/0.333
Kosmos-2	0.255/0.281	0.644/0.641	0.196/0.195	0.358/0.368	0.210/0.225	0.489/0.491	0.359/0.365	0.359/0.367
InstructBLIP (Flan-T5-XL)	0.334/0.362	0.582/0.599	0.248/0.267	0.113/0.113	0.167/0.188	0.378/0.400	0.211/0.179	0.290/0.301
InstructBLIP (Vicuna-7B)	0.359/0.437	0.683/0.689	0.200/0.283	0.253/0.367	0.263/0.304	0.629/0.663	0.337/0.382	0.389/0.446
VisualGLM-6B (GLM-6B)	0.247/0.234	0.498/0.507	0.146/0.154	0.110/0.116	0.209/0.183	0.342/0.349	0.127/0.131	0.240/0.239
mPLUG-Owl (LLaMA-7B)	0.409/0.427	0.634/0.644	0.241/0.271	0.437/0.487	0.148/0.180	0.687/0.711	0.466/0.486	0.432/0.458
LLaMA-Adapter-V2	0.354/0.363	0.464/0.506	0.275/0.329	0.298/0.360	0.257/0.271	0.604/0.666	0.412/0.425	0.381/0.417
LLaVA-v1 (Vicuna-13B)	0.462/0.457	0.442/0.462	0.264/0.280	0.404/0.417	0.208/0.237	0.626/0.684	0.349/0.372	0.394/0.416
MiniGPT-4 (Vicuna-13B)	0.239/0.257	0.238/0.253	0.170/0.183	0.339/0.340	0.252/0.246	0.572/0.591	0.239/0.233	0.293/0.300
Q-Instruct (Vicuna-v1.5-7B)	0.809/0.852	0.880/0.883	0.377/0.436	0.800/0.806	0.521/0.535	0.724/**0.828**	0.688/0.695	0.686/0.719
PhotoEye (Vicuna-v1.5-7B)	**0.823/0.857**	**0.891/0.897**	**0.407/0.478**	**0.824/0.822**	**0.531/0.544**	**0.778**/0.824	**0.701/0.706**	**0.708 / 0.733**

Table 2. Image Quality Assessment Results. Best results in **bold**.

ple. When all vision encoders are activated, PhotoEye outputs accurate and informative photo critiques that align well with critiques from photography professionals. When we deactivate CLIP, the generated critique mainly talks about the composition and overlooks the problem in lighting and exposure. By contrast, when we deactivate CoDETR, the model focuses on exposure again but is a little bit repetitive in last few sentences.

G. Evaluation on Image Quality Assessment

Because our PhotoCritique dataset mainly focuses on explainable visual aesthetic understanding with MLLM in real-world scenarios, i.e., aesthetic description, conversa-

Method	M=4	M=8	M=12	L=2	L=3	L=4	w/o L. Query	w/o M. Gating	w/o Fusor
Accuracy	72.68	**73.92**	73.91	72.56	**73.92**	**73.92**	71.26 **-2.66**	71.94 **-1.98**	68.83 **-5.09**

Table 3. **Blue** shows gap to Full Method. **Left:** Ablations on query quantity. **Middle:** Ablations on fusion block quantity. **Right:** Ablations on learnable queries (L. Query), multimodal gating (M. Gating), and *entire* vision fusor (Fusor).

Ablated Encoder	CLIP-ViT-L/14	DINOv2-giant	CoDETR-ViT-L	SAM-ViT-H	Full Method
Accuracy	69.90 **-4.02**	71.77 **-2.15**	72.74 **-1.18**	72.86 **-1.06**	**73.92**

Table 4. Vision Encoder Ablation. **Blue** shows gap to Full Method.

tion and VQA, our dataset does not contain traditional Image Quality Assessment (IQA) related training samples as Q-Instruct [3] does. To validate PhotoEye's generalizability to other tasks, we still evaluated PhotoEye with on IQA tasks. Tab. 2 demonstrates the effectiveness of PhotoEye on IQA tasks, where we also achieve competing results over baselines.

H. More Ablations

We have included the ablation of our entire vision fusor, the effect of different vision encoders and we show qualitative results in Fig. 7. In Tab. 3, we further provide detailed ablations on PhotoBench on: the number of learnable queries M, the number of fusion blocks L, learnable query module and multimodal gating. We validated our choices of M (8) and L (3), and show contributions of sub-modules in vision fusor. In Tab. 4, we show the performance drop when ablating each vision encoder.

I. Handling of Mixed Reviews

Aesthetics is subjective, and photographers may hold conflicting views about the attributes of an image. It is the uniqueness of aesthetic understanding: aesthetics sometimes do not have a definitive"yes-or-no" answer, as agreed by renowned critic John Berger. And even within prestigious groups like Magnum Photos, photographers can disagree over the evaluation of a single work. What really matters in such cases is not a "yes-or-no" judgment, but the justification behind the photographer's judgment. *The justifications reflect the essence of aesthetic understanding.* In our LLM-based photo critique generation, we aim to **preserve conflicting views if the corresponding *justification* is provided. When a particular opinion is dominant across all comments, it will be the primary opinion of the generated critique, while others are also mentioned.**

This approach results in more informative critiques that align with views of critics like Berger in *Ways of Seeing*. It also offers users more diverse perspectives to think and refine their photography skills or develop post-editing strategies in practice. Fig 1 in Appendix offers an example. There are conflicting views on the existence of the rainbow in comments. While #1, #3, #5, #6 love the rainbow for its exploding color, harmony with hill lights, or its number (double rainbow), #4 likes to remove it with the *justification* of it being blown out. Therefore while the generated critique primarily affirms the rainbow's presence (major), it also notes the concern on its blown-out look.

References

[1] Yipo Huang, Xiangfei Sheng, Zhichao Yang, Quan Yuan, Zhichao Duan, Pengfei Chen, Leida Li, Weisi Lin, and Guangming Shi. AesExpert: Towards multi-modality foundation model for image aesthetics perception. *arXiv preprint arXiv:2404.09624*, 2024. 2, 3

[2] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 3

[3] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, Geng Xue, Wenxiu Sun, Qiong Yan, and Weisi Lin. Q-Instruct: Improving low-level visual abilities for multi-modality foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 7

[4] Zhipeng Zhong, Fei Zhou, and Guoping Qiu. Aesthetically relevant image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2023. 1

[5] Zhaokun Zhou, Yiwei Su, Amin Zheng, Qiulin Wang, Rui Chen, Li Yuan, and Di Zhang. UNIAA: A unified multi-modal image aesthetic assessment baseline and benchmark. *arXiv preprint arXiv:2404.09619*, 2024. 3