



ECCV 2024

Depicting Beyond Scores: Advancing Image Quality Assessment through Multi-modal Language Models

Zhiyuan You (*Presenter*) · Zheyuan Li · Jinjin Gu · Zhenfei Yin · Tianfan Xue · Chao Dong

Project Page: <https://depictqa.github.io>

Evaluating the quality of images

- **With reference:**

- Some datasets (e.g., restoration) usually have a ground truth high-quality image
- PSNR, MSE, SSIM, LPIPS



Input



Output

Distance metric (i.e., a score) \updownarrow

- **No reference:**

- Some enhancement (like style transfer) or image generation, may not have a ground truth
- NIQE, MUSIQ, MANIQA



Ground-truth reference

Evaluating the quality of images

- **With reference:**

- Some datasets (e.g., restoration) usually have a ground truth high-quality image
- PSNR, MSE, SSIM, LPIPS

- **No reference:**

- Some enhancement (like style transfer) or image generation, may not have a ground truth
- NIQE, MUSIQ, MANIQA



Input

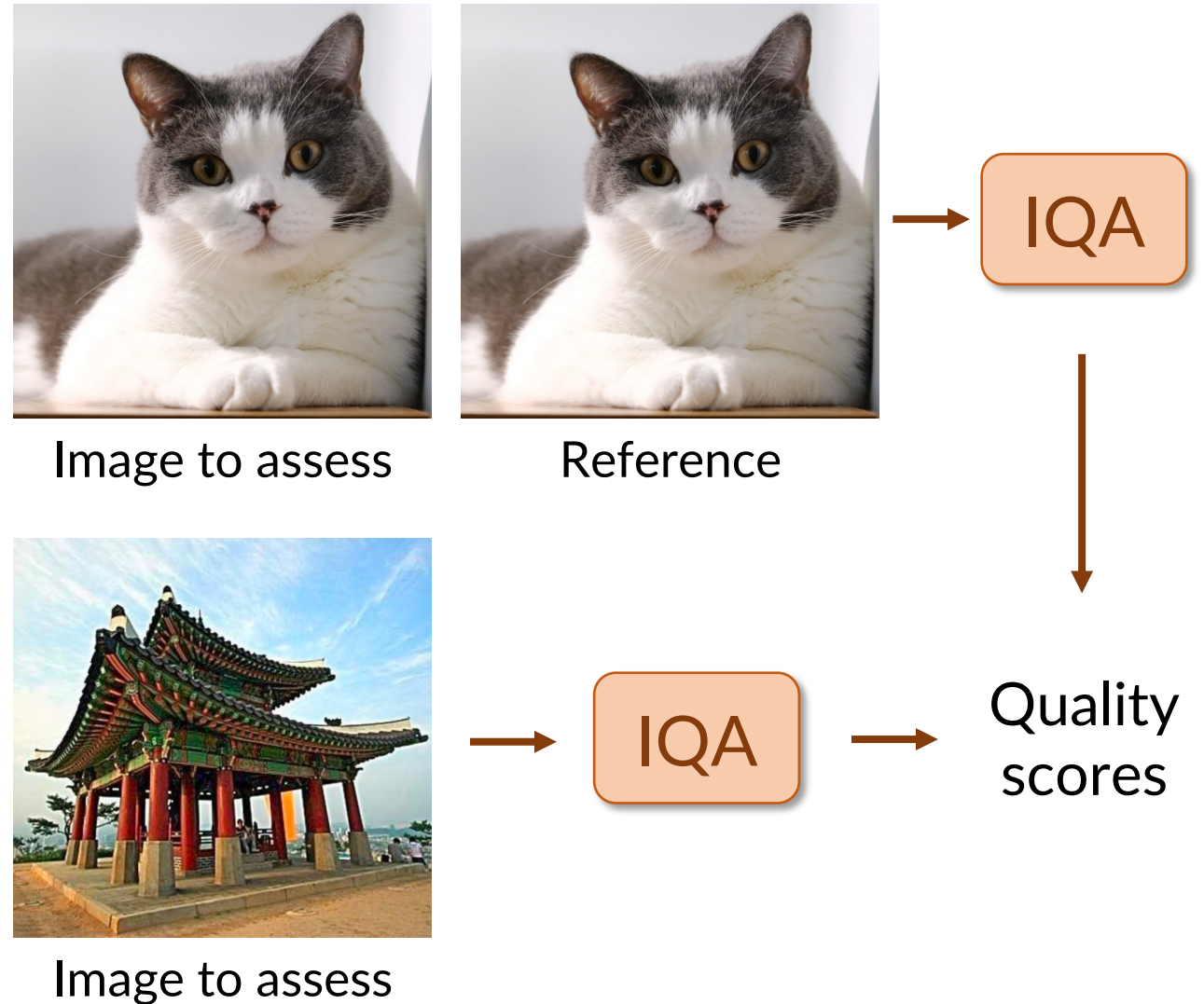


Output

↑
Model to
predict one
quality score

Evaluating the quality of images

- **With reference:**
 - PSNR, MSE, SSIM, LPIPS
- **No reference:**
 - NIQE, MUSIQ, MANIQA
- **However, is a single score enough?**



E.g., we all know noise is bad



MaskedDenoising

Chen et al., Masked Image Training for Generalizable Deep Image Denoising. CVPR 2023.

But, which images looks more realistic?



MUSIQ score ↑ 76.2

70.7

We simply add grain to the left image.

Actually, noise is not always bad. Minor noise may improve realistic.

<https://medium.com/photo-dojo/dont-fear-the-grain-263a37a64b87>

Is quality score reasonable?

- A simple score cannot reflect multiple aspects: noise, color, blurriness, etc.

Reference

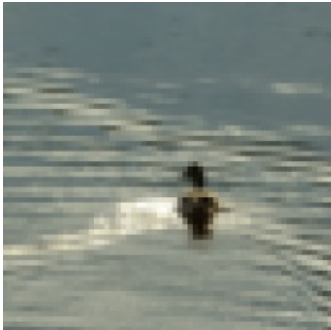
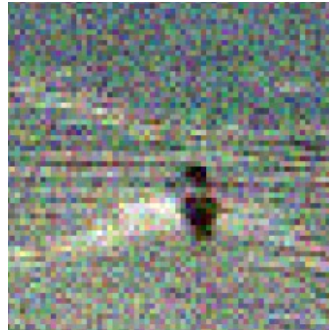
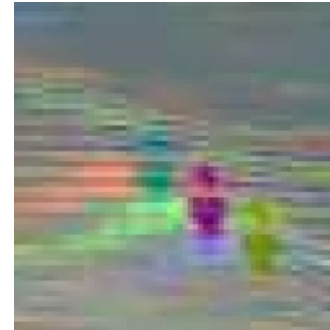


Image A



“Image is noisy, but content is still visible”

Image B



“Image has color artifacts, and content is hard to recognize”

By human

	Img.A	Img.B
PSNR↑	19.03	21.45 ✓
SSIM↑	0.37	0.67 ✓
LPIPS↓	0.45	0.29 ✓
PieAPP↓	1.83	1.22 ✓
Humans	✓	

By score

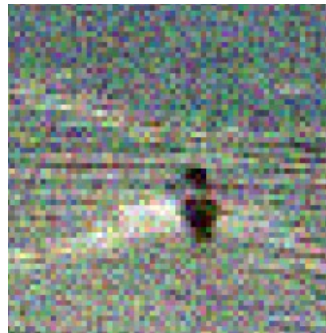
Humans do not use score, but

- Identify distortions
- Measure their impacts
- Draw conclusion

Reference

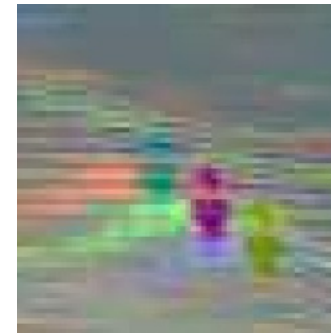


Image A



“Image is noisy, but content is still visible”

Image B



“Image has color artifacts, and content is hard to recognize”

By human

Can we mimic human using LLM?

- Identify distortions
- Measure its impact
- Draw conclusion

By our DepictQA

Reference



Image A

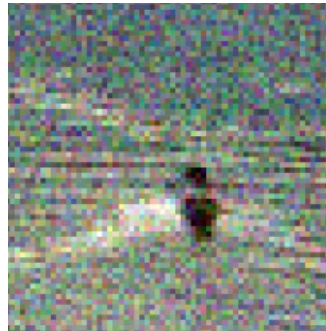
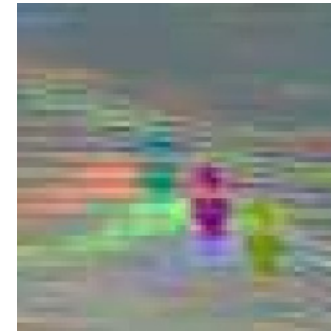


Image B



“Although Image A slightly outperforms Image B in terms of **color distortion and artifacts**, it falls short in the aspect of **noise**. Despite this, Image A still demonstrates a slight advantage over Image B in terms of overall distortion. Regarding the degree of texture damage, both images exhibit similar performance, **with the water surface texture barely recognizable**. Therefore, despite Image A's inferior performance in noise, its advantages in color distortion and artifacts contribute to **its slight superiority over Image B.**”




By human

“Image is noisy, but content is still visible”


“Image has color artifacts, and content is hard to recognize”

Our solution: DepictQA

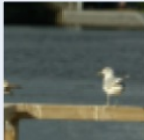
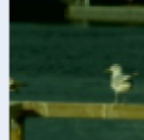


- **Our objectives:**
 - Mimics human IQA
 - Can handle different inputs

Reference:  Image:  


Analyze the image's quality, focusing on its key elements, and detail your findings.

 The overall quality of this image is **severely compromised**, primarily due to **the issue of noise**. Although the image maintains **good consistency in terms of brightness and color**, and the problems of **artifacts and blurring are well controlled**, the presence of noise significantly reduces the overall image quality. In terms of texture, **the architectural texture in this image can barely be recognized**, which also indicates a high degree of noise pollution. Therefore, despite the good performance in other aspects, the impact of noise overrides the advantages of other factors, leading to a serious degradation in overall quality and texture quality.

Single image assessment + optional reference

Reference:  Image A:  Image B:  

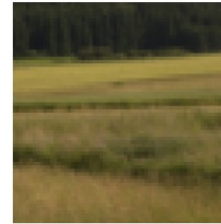
Assess the quality of Image A in relation to Image B, and provide a thorough comparison rationale.

 I am considering five key factors: brightness distortion, color distortion, noise, artifacts, and blurring. The results indicate that **Image A significantly outperforms Image B in terms of noise handling, but slightly underperforms in color distortion**. For brightness distortion, artifacts, and blurring, both images exhibit similar performance. Overall, the total distortion level of Image A is notably lower than that of Image B. Regarding the degree of texture damage, **the edge textures of objects in Image A can be barely identified, while those in Image B are completely damaged**. This may be due to the impact of noise, as it significantly degrades the texture quality of Image B. Therefore, based on the above analysis, I conclude that **the quality of Image A is superior to that of Image B**. Although Image A slightly falls short in color distortion handling, its advantage in noise control results in a lower overall distortion level and better texture quality.

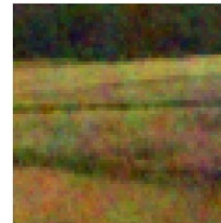
Paired image comparison + optional reference

Task-1: quality description

- **Human:** assess image quality by identifying **distortions & texture damages**.
- **Input:** a reference (optional) and a distorted image.
- **Task :** describe distortions and texture damages, then comment on overall quality.



Reference
Image



Distorted
Image

Resonse for Quality Description

The overall quality of this image has been **severely compromised**, particularly in **color distortion and noise**, resulting in a significant deviation in the overall visual perception of the image, despite maintaining good consistency in terms of brightness, artifacts, and blurring. In terms of texture quality, **the plant textures are barely recognizable**, primarily due to the influence of color distortion and noise. Therefore, the severe distortion in color and noise has **significantly affected** the overall and texture quality of the image.

Task-2: quality comparison

- **Human:** reliable on **comparing 2 images**. (single-image quality rating is hard!)
- **Input:** a reference (optional) and 2 distorted images.
- **Task:** determine which one is better.



Reference



Image A



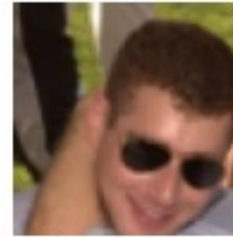
Image B

Response for Quality Comparison

Image A maintains a distinct advantage in terms of image quality over Image B, substantiating its superior qualities.

Task-3: comparison reasoning

- **Human:** compare 2 images and **sharing reasoning**.
- **Input:** a reference (optional) and 2 distorted images.
- **Task:** compare by identifying distortions and weighing the pros and cons.



Reference

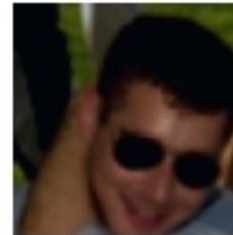


Image A

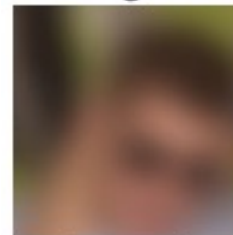


Image B

Response for Comparison Reasoning

Although Image A is **slightly inferior** to Image B in terms of **brightness distortion**, it **significantly outperforms in blurriness**. From the perspective of overall distortion, Image A far outperforms Image B. Image A can **clearly identify the facial textures**, whereas the facial textures in Image B are **completely damaged** and unidentifiable. Therefore, despite Image A's slight inferiority in brightness distortion, its advantages in blurriness and its ability to clearly identify textures, result in **Image A's overall quality being superior**.

Dataset creation, source 1: detailed data

- Human annotates for Task-1 & Task-3.
- Detailed & high-quality
- But small amount (5k)



Reference Image



Distorted Image

Texture	Distortion	Overall Quality
Texture Category: <input type="checkbox"/> Water Surface <input type="checkbox"/> Fabrics <input checked="" type="checkbox"/> Plants, Foliage <input type="checkbox"/> Faces	Brightness Distortion: <input checked="" type="checkbox"/> Undistorted Mild Severe Color Distortion: <input type="checkbox"/> Undistorted Mild <input checked="" type="checkbox"/> Severe Noise: <input type="checkbox"/> Undistorted Mild <input checked="" type="checkbox"/> Severe Artifacts: <input checked="" type="checkbox"/> Undistorted Mild Severe Blurriness: <input checked="" type="checkbox"/> Undistorted Mild Severe	Distorted Image is: <input type="checkbox"/> Nearly undistorted <input type="checkbox"/> Mild distortion <input checked="" type="checkbox"/> Severe distortion  Annotators Complete Questionnaire

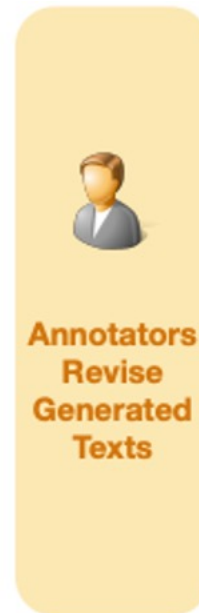
Human Annotate



GPT-4 Generate



Human Revise



Reason for Quality Description

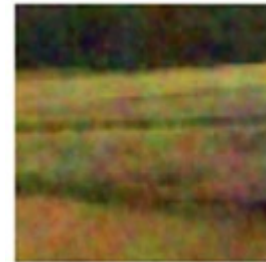
The overall quality of this image has been **severely compromised**, particularly in **color distortion and noise**, resulting in a significant deviation in the overall visual perception of the image, despite maintaining good consistency in terms of brightness, artifacts, and blurring. In terms of texture quality, **the plant textures are barely recognizable**, primarily due to the influence of color distortion and noise. Therefore, the severe distortion in color and noise has **significantly affected** the overall and texture quality of the image.

Dataset creation, source 2: templated data

- Templated data for Task-2 (quality comparison).
 - Which one is better: constructed from existing datasets
 - **Templated & brief**
 - But **large amount** (125k)



(is better than)



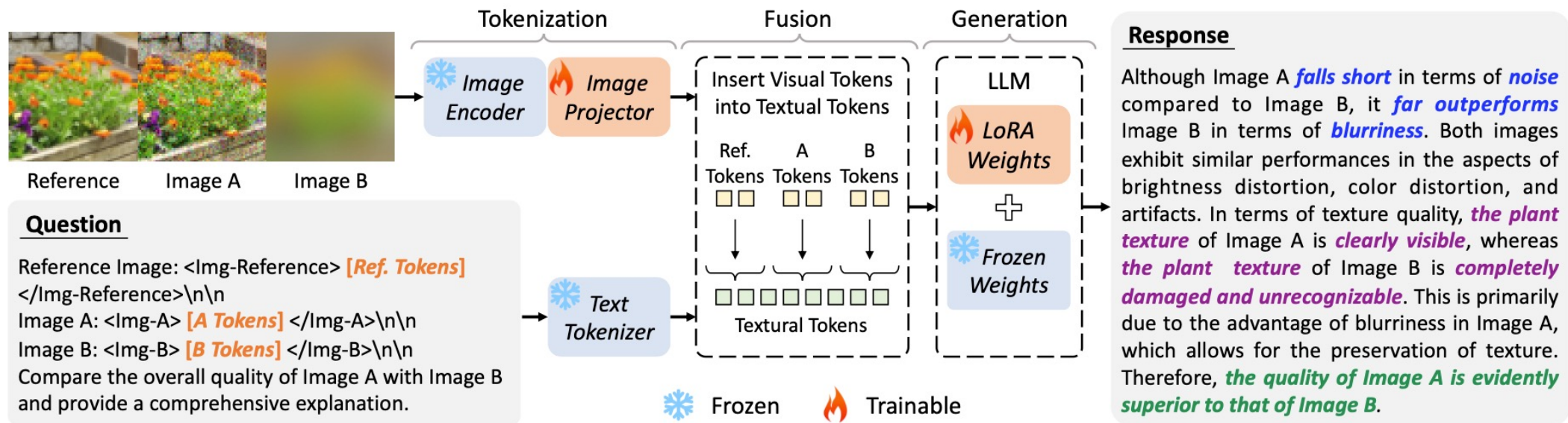
Sample from GPT-4
Generated One-Sentence
Comparative Conclusion

Response for Quality Comparison

Image A maintains a distinct advantage in terms of image quality over Image B, substantiating its superior qualities.

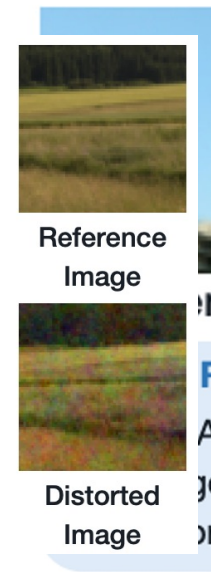
Model architecture

- **Visual encoder:** CLIP pre-trained ViT-L/14.
- **Text tokenizer:** SentencePiece tokenizer.
- **Token fusion:** insert visual tokens into textual tokens.
- **LLM fine-tuning:** LoRA.



Training with multi-source data

- **Data 1:** 115,646 brief **templated** data (quality comparison).
 - **Effects:** The abundant templated data principally bridge images and descriptive texts in quality-related tasks.
- **Data 2:** 4,854 **detailed** data (quality description and comparison reasoning).
 - **Effects:** The limited yet high-quality data foster the model's descriptive and reasoning skills.
- **Data 3:** 48,734 **external** content description data (MS-COCO), unrelated to IQA.
 - **Effects:** The IQA-unrelated content description data serve as regularization, given the limited text diversity of IQA data for an MLLM.



Resonse for Quality Description

The overall quality of this image has been **severely compromised**, particularly in **color distortion and noise**, resulting in a significant deviation in the overall visual perception of the image, despite maintaining good consistency in terms of brightness, artifacts, and blurring. In terms of texture quality, **the plant textures are barely recognizable**, primarily due to the influence of color distortion and noise. Therefore, the severe distortion in color and noise has **significantly affected** the overall and texture quality of the image.

How training data helps

Templated data 1 (115,646): pre-trains the basic weights for IQA



Reference Image A Image B

Response for Quality Comparison

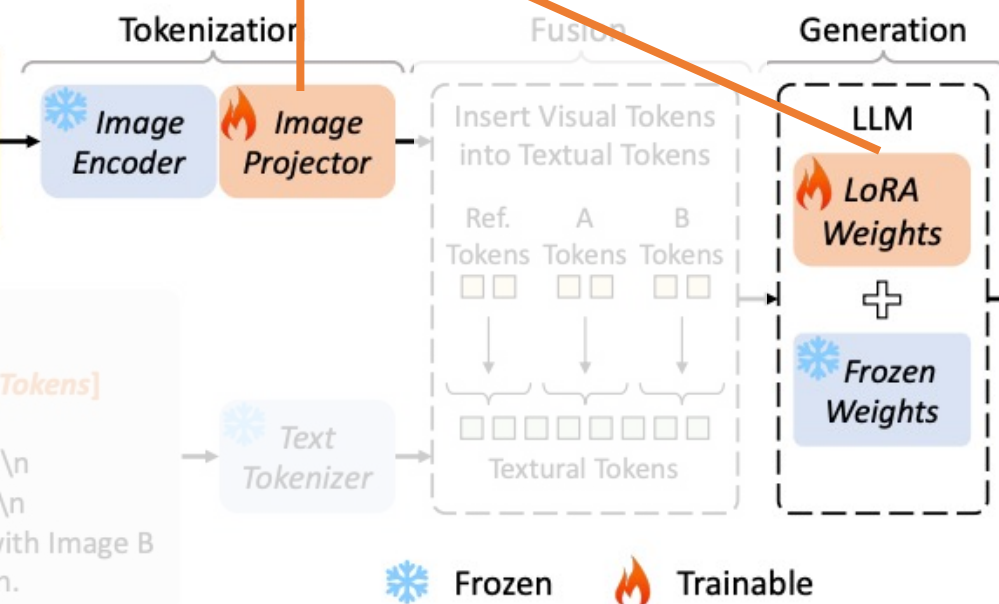
Image A maintains a distinct advantage in terms of image quality over Image B, substantiating its superior qualities.



Reference Image A Image B

Question

Reference Image: <Img-Reference> [Ref. Tokens] </Img-Reference>\n\n
 Image A: <Img-A> [A Tokens] </Img-A>\n\n
 Image B: <Img-B> [B Tokens] </Img-B>\n\n
 Compare the overall quality of Image A with Image B and provide a comprehensive explanation.



Response

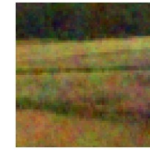
Although Image A *falls short* in terms of *noise* compared to Image B, it *far outperforms* Image B in terms of *blurriness*. Both images exhibit similar performances in the aspects of brightness distortion, color distortion, and artifacts. In terms of texture quality, *the plant texture* of Image A is *clearly visible*, whereas *the plant texture* of Image B is *completely damaged and unrecognizable*. This is primarily due to the advantage of blurriness in Image A, which allows for the preservation of texture. Therefore, *the quality of Image A is evidently superior to that of Image B*.

How training data helps

Detailed data 2 (4,854): A small set of detailed descriptive data improves the descriptive and reasoning skill.



Reference Image



Distorted Image

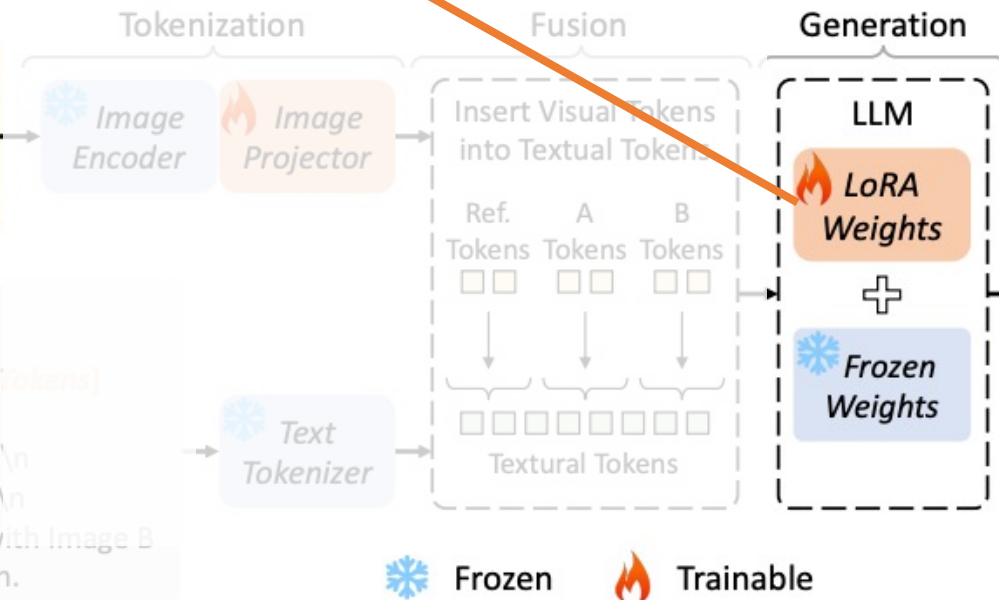
Response for Quality Description

The overall quality of this image has been **severely compromised**, particularly in **color distortion and noise**, resulting in a significant deviation in the overall visual perception of the image, despite maintaining good consistency in terms of brightness, artifacts, and blurring. In terms of texture quality, **the plant textures are barely recognizable**, primarily due to the influence of color distortion and noise. Therefore, the severe distortion in color and noise has **significantly affected** the overall and texture quality of the image.



Question

Reference Image: <Img-Reference> [Ref. Tokens] </Img-Reference>\n\nImage A: <Img-A> [A Tokens] </Img-A>\n\nImage B: <Img-B> [B Tokens] </Img-B>\n\nCompare the overall quality of Image A with Image B and provide a comprehensive explanation.

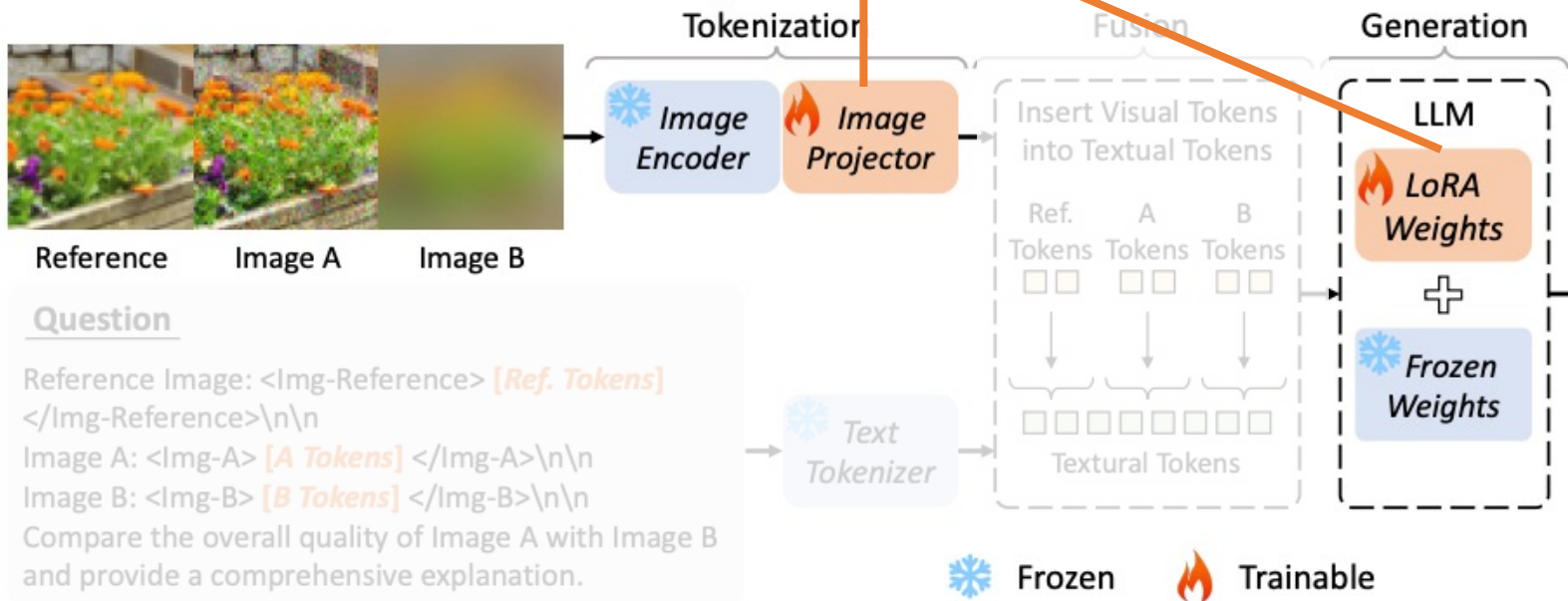


Response

Although Image A **falls short** in terms of **noise** compared to Image B, it **far outperforms** Image B in terms of **blurriness**. Both images exhibit similar performances in the aspects of brightness distortion, color distortion, and artifacts. In terms of texture quality, **the plant texture** of Image A is **clearly visible**, whereas **the plant texture** of Image B is **completely damaged and unrecognizable**. This is primarily due to the advantage of blurriness in Image A, which allows for the preservation of texture. Therefore, **the quality of Image A is evidently superior to that of Image B**.

How training data helps

Non-IQA data 3 (48,734): regularizes the LLM to prevent repeated words and limited vocabulary



Question

Reference Image: <Img-Reference> [Ref. Tokens] </Img-Reference>\n\nImage A: <Img-A> [A Tokens] </Img-A>\n\nImage B: <Img-B> [B Tokens] </Img-B>\n\nCompare the overall quality of Image A with Image B and provide a comprehensive explanation.

Response

Although Image A *falls short* in terms of *noise* compared to Image B, it *far outperforms* Image B in terms of *blurriness*. Both images exhibit similar performances in the aspects of brightness distortion, color distortion, and artifacts. In terms of texture quality, *the plant texture* of Image A is *clearly visible*, whereas *the plant texture* of Image B is *completely damaged and unrecognizable*. This is primarily due to the advantage of blurriness in Image A, which allows for the preservation of texture. Therefore, *the quality of Image A is evidently superior to that of Image B*.

Results

Pair-wise comparison on BAPPS
Compare to score-based IQA

Type	Method	Comparison
Oracle	Human	80.8 / 84.4
Traditional	PSNR	59.9 / 77.8
	SSIM [61]	60.3 / 79.1
	VIF [52]	55.6 / 74.4
	FSIM [72]	62.7 / 79.4
Learning	DeepIQA [4]	70.3 / 79.4
	PieAPP [47]	72.7 / 77.0
	LPIPS [75]	76.0 / 82.8
	DISTS [15]	77.2 / 82.2
DepictQA (Ours)		80.3 / 84.2

Method	LLaVA [30]	LLaVA-1.5 [29]	MiniGPT4 [82]	MiniGPT4 [82]	LAMM [69]	LAMM [69]	DepictQA (Ours)
LLM (7B)	LLaMA-2c	Vicuna-v1.5	Vicuna-v0	LLaMA-2c	Vicuna-v0	LLaMA-2c	Vicuna-v1.5
Description	fail	18.0 / 65.8	16.0 / 49.9	fail	12.0 / 62.5	8.0 / 57.4	64.0 / 76.2
Comparison	fail	43.0 / 50.6	38.0 / 46.3	fail	53.0 / 55.8	44.0 / 48.9	82.0
Reasoning	fail	7.0 / 63.9	1.0 / 42.4	fail	4.0 / 58.1	5.0 / 52.4	53.0 / 76.4

Detailed descriptive ability
Compare to multi-modal LLMs

Quality comparison results

Reference:

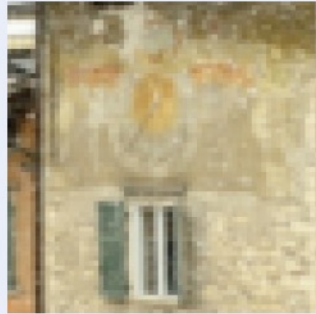


Image A:

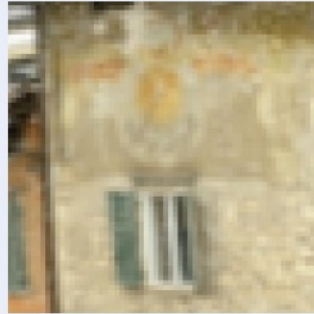
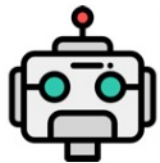


Image B:



Make a judgment on which image, Image A or Image B, you consider to be of better quality.



I consider **Image A** to be of better quality.

(a)

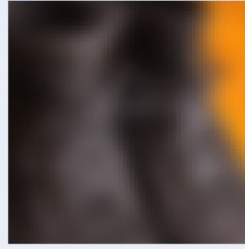
Human labeled ground-truth: **Image A**

Quality description results

Reference:



Image:





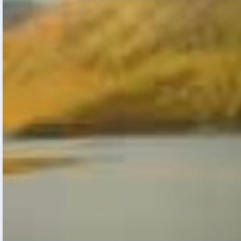
In your opinion, what are the strengths and weaknesses of the image in terms of quality, and why?



The overall quality of this image is **severely compromised**, particularly in terms of **blurriness**, which has resulted in significant distortion. Although the image maintains a **good quality in terms of brightness, color, noise, and artifacts**, with no apparent distortion, the severe distortion caused by blurriness has greatly reduced the overall image quality. Additionally, from a texture perspective, **the texture of this image is completely damaged and unrecognizable**, which further demonstrates the severe impact of blurriness on image quality. Therefore, it can be concluded that the main issue with this image lies in its blurriness, which is also the primary cause of its overall decline in quality and texture damage.

(c)

Comparison reasoning results

Reference:  , Image A:  , Image B:  ,

How would you rate Image A and Image B in terms of quality, and what specific aspects contribute to their varying levels of quality?



Image B exhibits superior quality compared to Image A.

Although both images perform similarly in terms of brightness distortion and noise, **Image A falls short in color distortion and artifacts compared to Image B.** Additionally, **Image A slightly outperforms in terms of blurriness**, but this does not compensate for its deficiencies in other areas. From the perspective of overall distortion, the quality of Image A is also noticeably lower than that of Image B. **In terms of texture degradation, both images can barely identify the edges of objects, indicating that factors such as color distortion and artifacts have caused significant damage to the texture.** Therefore, considering all these factors, the quality of Image B is superior to that of Image A.

(a)

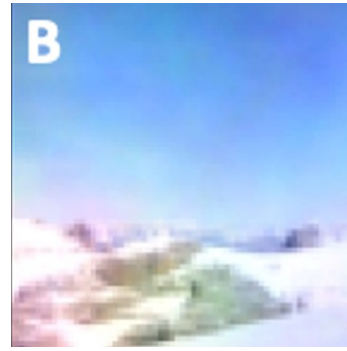
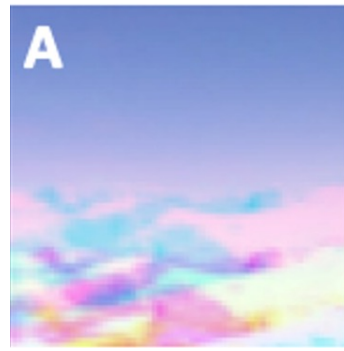
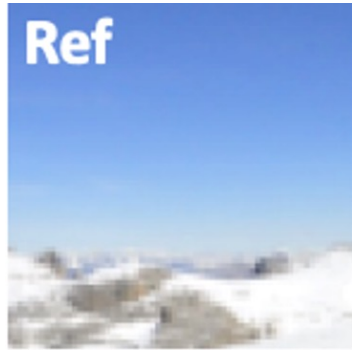
Extension: no reference extension

Task	Performance
Description	76.4 / 84.5
Comparison	82.3 / 78.5
Reasoning	77.2 / 78.8

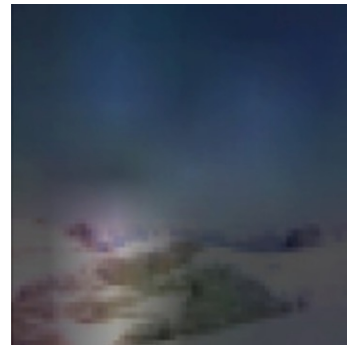
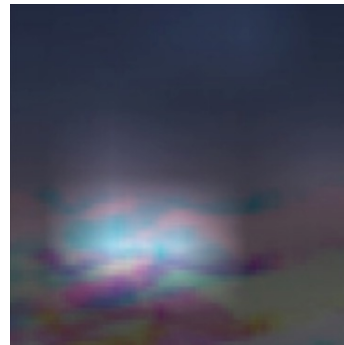
With reference / no reference performance

Although gathered in a full-reference setting, our dataset can be used in non-reference applications.

Extension: distortion localization



“...The quality of Image B is significantly superior to that of Image A. Image A noticeably falls short in two key factors, **color distortion and artifacts...**”



Back propagate these words to the image

Color distortion is here

(This part is still work-in-progress and may not always work)

More are coming!

- **DepictQA-v2**

- More comprehensive (**3→8**) **task paradigm**
- Much larger scale (**10×**) **datasets**
- More comprehensive (**3×**) **distortions**
- More efficient **dataset construction** with GPT-**4V** (v.s., DepictQA-v1 uses GPT-4)
- Support **flexible image resolutions** and **confidence estimation**
- From toy dataset (BAPPS in DepictQA-v1) to **real-world images**

You et. al., Descriptive Image Quality Assessment in the Wild. arXiv 2024.

<https://depictqa.github.io>

DepictQA-v2 assessing web-downloaded images



The image depicts a close-up of a pink flower with green foliage in the background. The image suffers from **significant blur**, which **obscures fine details such as the texture of the petals and the structure of the flower's center**. Additionally, there is a **slight compression artifact** present, which can be observed as a **subtle blockiness**, particularly noticeable in areas of subtle color gradation. Overall, the image quality is **considerably degraded** due to the severe blur. The image lacks clarity and sharpness, making it difficult to appreciate the intricate details of the flower.



The image depicts a sunset with a vibrant orange sky above a silhouette of a landscape. The image exhibits **slight compression** and **obvious pixelation**. The compression has led to a loss of fine details, with **blocky artifacts visible**, especially in areas of **subtle color gradients**. The severe pixelation results in a significant loss of detail, which makes **the silhouette of the landscape and the sunset's gradient appear as large, flat areas**. Overall, the quality is **substantially degraded**, with the pixelation being most serious, as it severely impacts the natural gradation of the scene.

You et. al., Descriptive Image Quality Assessment in the Wild. arXiv 2024.

<https://depictqa.github.io>



Thanks