

R4-CGQA: Retrieval-based Vision Language Models for Computer Graphics Image Quality Assessment

1. Introduction

Our work provides an indicator for CG development, highlighting the key dimensions in the CG rendering process. It also indicates that a general large visual language model can effectively obtain knowledge from similar images, providing important evidence for developing AI vision. In Section 2, we give the details about dataset dimension. In Section 2.1, we introduce some prompts. In Section 3, we perform experiments on LlavaNext to show introducing bad (irrelevant) examples will decrease performance on CGQA. In Section 4, we visualize some results of our system for comparison.

2. Dataset information

As shown in Table 1, we give an explanation of these dimensions. The table defines the six quality dimensions we use throughout the paper to analyze and explain CG image quality. The first group **Lighting**, **Material**, and **Color** captures appearance- and photometry-level factors (illumination adequacy, surface properties, tonal balance). The second group **Atmosphere**, **Realism**, and **Space** captures scene- and perception-level factors (mood / aesthetics, physical plausibility, spatial composition and depth). These dimensions are *complementary rather than mutually exclusive*: a single CG may be strong/weak on several axes simultaneously. We use the same vocabulary when (i) writing reference descriptions for retrieval, (ii) constructing prompts for the VLM, and (iii) interpreting answers, which improves consistency and interpretability of our evaluation.

As shown in Figure 1, The style distribution of a sampled subset (n=300) shows that the dataset is dominated by realistic images, with Modern Realistic (36.3%) and Medieval Realistic (16.3%) being the two largest categories. At the same time, the dataset also covers multiple stylized genres, including Western Cartoon (12.3%), Realistic (General) (9.0%), Sci-Fi (7.3%), Fantasy (7.0%), and Japanese Anime (5.7%), indicating good style diversity. The image resolutions are highly diverse. As shown in the width-height scatter plot, most samples are concentrated in the medium-to-high resolution range, while a few images reach very high resolutions, which increases the visual diversity of the

Table 1. Explanations of 6 key CG quality dimensions

Dimension	Explanation
Lighting	Primarily evaluates whether the lighting is appropriate and soft, the presence of overexposure or dim illumination, the naturalness of shadow transitions, and the clarity of the light source.
Material	Focuses on whether the surface texture and material details of objects are distinct and consistent with their real properties, including structure, texture, glossiness, and transparency.
Color	Assesses the richness and subtlety of colors, the ability to create harmonious warm or cool tones, and whether saturation and contrast levels are sufficient and appropriate.
Atmosphere	Concerns the overall mood and aesthetic impression of the image, such as whether it conveys tranquility or dynamism, warmth or coldness, dreaminess or realism, mystery or clarity, and whether it evokes artistic beauty or emotional impact.
Realism	Evaluates the extent to which the image conforms to physical and natural laws, the degree of photorealism, and the viewer’s sense of presence, immersion, and believability.
Space	Examines background clarity and subject prominence, as well as the naturalness of the transitions between objects and background in terms of texture and material blending, ensuring coherent composition and depth hierarchy.

dataset. The score distribution is mainly concentrated on scores 3 and 4, with 1162 and 1260 samples, respectively. In comparison, very few samples are assigned score 0 or 1, suggesting that the dataset is dominated by medium- to high-quality images while still retaining a certain amount of low-quality cases.

2.1. Prompt suite overview

We employ a compact set of prompts to (i) *generate* question-answer items from human-written CG quality descriptions, (ii) *test* VLMs under retrieval-augmented conditions, and (iii) *score* free-form answers automatically with ChatGPT-4o. All prompts are aligned with the six quality

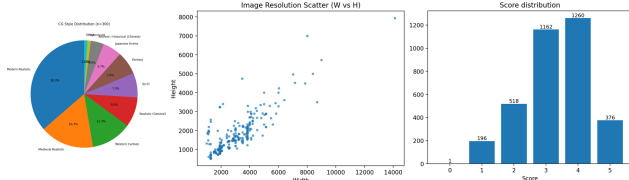


Figure 1. More quantitative information of our dataset

dimensions (see Table 1), i.e., lighting, color, material, atmosphere, realism, and space, and enforce concise outputs in JSON for deterministic parsing and reproducibility.

Generation prompts. We use ChatGPT-4o to synthesize three types of items from a given description and its overall score: (1) *Choice generation* creates multi-choice questions with one correct answer and three plausible but conflicting distractors (false candidates). (2) *Yes-or-no generation* produces balanced binary questions (roughly equal “Yes”/“No”) and allows asking about potential issues not explicitly mentioned (answered with “No” if uncertain). (3) *Q&A generation* yields five concise open questions with short answers. All outputs are lists in JSON only.

Test prompts. For evaluation, each query image is accompanied by a retrieval-selected reference description (our dual-embedding module) and a task-specific instruction: (1) *Choice test* requires returning a single letter (A–D). (2) *Yes-or-no test* requires “Yes” or “No” only. (3) *Q&A test* asks for a brief free-form answer. This setup probes the VLM both on discrete decisions and short explanations while keeping outputs machine-checkable.

Scoring prompt. An impartial ChatGPT-4o rater compares a model answer with the ground truth and returns a single integer in [1, 5] (5 = identical), enabling automatic, fine-grained assessment without training or tuning the VLM.

2.2. Prompt for choice generation by ChatGPT-4o

Generate multiple question and answer pairs based on the following description of an CG image’s quality, considering aspects of light, color, material quality, atmospheric quality, realism, spatial and overall quality. A overall quality score will be given (1 is the worst, 5 is the best). The questions can begin with ‘What/Why/Where/How’. The answer should be concise, containing only the essential information in minimal words. For each question, also produce three false answers under the key ‘false candidates’ that are plausible but conflict with the description. Format everything as a list in JSON. When you respond, please only output the JSON.

Description: < desc >

Score: < score >

2.3. Prompt for yes-or-no generation by ChatGPT-4o

Generate multiple yes-or-no question and answer pairs based on the following description of a CG’s quality, focusing on light, color, material quality, atmospheric quality, realism, spatial quality and overall quality. Besides, an overall quality score will be given (1 is the worst, 5 is the best). The answer should only contain ‘Yes’ or ‘No’, and the number of ‘Yes’ answers should be close to the number of ‘No’ answers. You may also ask questions about possible quality issues not mentioned in the description, but answer them with ‘No’ if uncertain Organize everything as a list in JSON. When you respond, please only output the JSON, with no additional text:

Description: < desc >

Score: < score >

2.4. Prompt for Q&A generation by ChatGPT-4o

Generate 5 question and answer pairs based on the following description of a CG image’s quality, considering aspects of light, color, material quality, atmospheric quality, realism, spatial or overall quality. An overall quality score will be given (1 is the worst, 5 is the best). The questions can begin with ‘What/Why/Where/How’. The answer should be concise, containing only the essential information in minimal words. Organize everything as a list in JSON. When you respond, please only output the JSON, with no additional text:

Description: < desc >

Score: < score >

2.5. Prompt for choice test

You will answer questions about the target image. A similar reference CG image has the following quality description:

< example desc >

Now look at the CG image and answer:

< Question >

A. < ChoiceA >

B. < ChoiceB >

C. < ChoiceC >

D. < ChoiceD >

Respond with ONE letter (A–D) only.

2.6. Prompt for yes-or-no test

You will answer questions about the target image. A similar reference CG image has the following quality description:

< example desc >

Now look at the CG image and answer:

Yes/No Question: < Question >

Respond with Yes or No.

2.7. Prompt for Q&A test

You will answer questions about the target CG image.
A similar reference example CG image has the following quality description:
< example desc >
Now look at the CG image and answer:
Question: < Question >

2.8. Prompt for ChatGPT-4o scoring

You are an impartial evaluator for CG image question-answering.
Given a question (which assumes the evaluator can see the image),
the ground-truth answer (GT), and the model’s answer (A),
rate how close A is to GT
on a 1–5 integer scale: 5 = identical / paraphrase
4 = minor differences that do not change meaning
3 = partially correct; key details missing
2 = mostly wrong but weak relevance
1 = completely wrong or irrelevant
Reply with a single integer only.
Compare the following two CG image quality answers.
Reflect how well the MODEL description matches the HUMAN description.
Output only a single score number between 1 (worst) and 5 (best).
HUMAN description:
< gt answer >
MODEL description:
< model answer >
Score:

3. Bad example decreases performance

As shown in Table 2, we compare five settings consistent with our Bayesian retrieval view. *Base* denotes no retrieval, *Bad example* (an irrelevant example is forced, effectively picking an index with very low posterior), On Choice questions, the full pipeline reaches 71.7%, outperforming both single-view variants (70.4% quality-only, 69.9% content-only) and the Base model (70.0%), while an irrelevant example hurts (69.1%). On Yes-or-no questions, the gains are larger: the full pipeline achieves 69.3% vs. 68.0% (quality-only), 66.3% (content-only), and 64.1% (Base); a bad example again degrades to 63.0%. These results support our design: (i) retrieval helps only when the selected index has high posterior relevance (the “Bad example” rows provide a counter-example); (ii) quality-aware neighbors are especially informative for binary decisions (quality-only > content-only by +1.7 points on Yes-or-no); and (iii) fusing content and quality yields the most robust improvements, consistent with the complementary likelihoods assumed in our MAP formulation.

Table 2. Results on LlaveNext-8B; Bad example denotes we give an irrelevant content as the example for VLM.

Setting	Bad example	Base	Without content	Without quality	Full pipeline
Choice	69.1%	70.0%	70.4%	69.9%	71.7%
Yes-or-no	63.0%	64.1%	68.0%	66.3%	69.3%

4. System visualization results

Qualitative comparison on four queries (Fig. 2 → Fig. 5). Across all four cases, our retrieval-augmented pipeline (R4-CGQA) produces answers that are (i) explicitly grounded in the six quality dimensions (lighting, material, color, atmosphere, realism, space), (ii) better calibrated—able to praise or criticize with reasons, (iii) more concise yet actionable, and (iv) supported by the retrieved neighbors (shown under each panel) with their CLIP / REIQA similarities.

Case 1 (war-zone street). *Original* gives a generic disclaimer about “game screenshots,” offering little evidence. *R4-CGQA* enumerates realism, detail, and overall aesthetics, and concludes with a justified score; the retrieved samples depict similar street battle scenes, anchoring the judgment.

Case 2 (toy bear in room). Both methods are positive, but *R4-CGQA* ties its conclusion to lighting, color harmony, facial expression, and playful atmosphere, producing a clearer rationale; retrieved frames from similar scenes further constrain the answer.

Case 3 (back-lit dragon). *Original* over-estimates the quality (“high-quality CG”). *R4-CGQA* identifies concrete defects—over-exposure, weak depth, lack of clarity—and downgrades the decision, showing that retrieval helps the model argue *against* quality when evidence is negative.

Case 4 (game character portrait). *Original* labels the image as low quality with broad statements. *R4-CGQA* yields a balanced “average” assessment with specific reasons (hair density, leather texture, lighting transitions), demonstrating finer granularity and dimension coverage; retrieved portraits with similar lighting support the critique. Overall, these visualizations show that our Bayesian RAG retrieval selects relevant exemplars that guide the VLM toward dimension-aware, evidence-backed explanations and more reliable final judgments.

Original

Q: How about the image quality, you can output a score 1-5, 5 is the best.

A: The image provided appears to be a screenshot from a video game. However, since it's a still image, it's not a typical photograph that would be subjected to an "image quality" assessment in the same way as a real photograph would be. The resolution and clarity within the context of the game seem consistent with the graphics standards for modern gaming platforms. If we were to evaluate its quality based on the criteria for images, it might score high in sharpness and contrast due to the detailed textures and vivid colors typical of contemporary video games. If you're asking about the content or the overall appearance within the context of the game, I would rate it as a 5 for visual appeal within its intended medium, which is a video game. However, if you're looking for real-world image quality metrics such as resolution, focus, and color accuracy, this screenshot doesn't provide enough information to accurately assess.

How about the image quality, you can output a score 1-5, 5 is the best.



R4-CGQA

Q: How about the image quality, you can output a score 1-5, 5 is the best.

A: As an AI visual assistant, I will evaluate this CG image based on its quality in terms of realism, attention to detail, and overall aesthetic appeal. While I am not an expert in CGI evaluation, I can provide you with a general score based on these criteria:

- 1. Realism:** The image shows a chaotic battle scene with a group of tanks firing and buildings in flames. The details of the damage and destruction are well-executed, creating a believable depiction of a war zone. A high score for realism.
- 2. Attention to Detail:** There is a lot of detail in the image, from the individual panels showing tank hits and fires to the smoke trails and debris. However, the image's overall coherence might be compromised by the number of elements competing for attention. A score that leans toward realism but may lose some points for clutter.
- 3. Overall Aesthetic Appeal:** The contrast between warm and cold tones in the image is striking, which enhances the dramatic tension of the scene. The lighting is also dynamic, with shadows and highlights adding depth to the environment. However, the material damages can be distracting, as they seem excessive and somewhat at odds with what would be experienced in a real battlefield. A score that is relatively high but could benefit from some refinement.




Considering these points, I would give this CG image a score of 4 out of 5 for quality, taking into account its realistic depiction, attention to detail, and overall aesthetic appeal. The image effectively conveys the chaos and intensity of a battlefield but could benefit from some refinement in the areas of material damage and perhaps a more focused narrative.

example: On

PASS (avg=0.953)

sim_clip=0.923, sim_reiqa=0.984

▼ View the retrieved candidate images (sorted by average similarity)

Thumbnails	Name	CLIP Sim	REIQA Sim	Average
	00393.png	0.923	0.984	0.953
	01568.png	0.929	0.978	0.953
	01100.png	0.918	0.974	0.946

How about the image quality, you can output a score 1-5, 5 is the best.



Figure 2. The actual measurement of question-answering dialogues and retrieval effects on the system; the VLM is Llava-7B.

Original

Q: What about the quality of this image?

A: The image is a computer-generated (CG) image of a pink teddy bear in a toy room, and the quality of the image is very good. The image is clear and detailed, with no visible artifacts or defects. The colors are vibrant and well-rendered, and the textures and lighting are realistic. The composition of the image is also well-balanced, with the teddy bear placed centrally in the frame and surrounded by toys and other objects. Overall, the image is a high-quality representation of a toy room, with a focus on the pink teddy bear as the main subject.

What about the quality of this image?



R4-CGQA

Q: What about the quality of this image?

A: The image you provided is a still from the movie "Toy Story 4," featuring the character Lotso as a pink bear with a smile on his face. The quality of the image is good, with clear details and vibrant colors. The character's facial expression is also well-rendered, with a subtle smile that adds to the overall charm of the image.

The background of the image is also well-done, with a colorful and lively environment that complements the character's personality. The overall atmosphere of the image is fun and playful, which is fitting for a character like Lotso who is known for his fun and mischievous personality.

Overall, the quality of the image is good, with clear details and a fun and playful atmosphere. The character's facial expression and the background environment are both well-done, making for a fun and engaging image.





I hope this helps to answer your question. Let me know if you have any other questions or if there's anything else I can help you with.

example: On

PASS (avg=0.866)

sim_clip=0.838, sim_reiqa=0.893

▼ View the retrieved candidate images (sorted by average similarity)

Thumbnails	Name	CLIP 相似度	REIQA 相似度	Average
	00872.jpg	0.838	0.893	0.866
	02966.jpg	0.830	0.901	0.865
	00669.jpg	0.820	0.874	0.847
	02291.jpg	0.819	0.858	0.839
	02719.jpg	0.840	0.767	0.803

What about the quality of this image?




Figure 3. The actual measurement of question-answering dialogues and retrieval effects on the system; the VLM is Llava-7B.

Original

Q: How about the quality of this CG image? is it has a good quality?

A: The CG image you've provided appears to have a high level of detail and quality. The textures on the dragon's scales are intricate and well-rendered, and the lighting and shading give the image a realistic and immersive feel. The overall composition and the level of detail suggest that this is a high-quality CG image, likely created with advanced rendering techniques and software.

How about the quality of this CG image? is it has a good quality?



R4-CGQA

Q: How about the quality of this CG image? is it has a good quality?




A: The quality of the user's CG image can be evaluated based on the reference description provided. Here are the key points to consider:

- Lighting and Exposure:** The image appears to have a high level of brightness, which might be intentional to create a certain atmosphere or effect. However, the overexposure and unnatural brightness changes can detract from the overall quality, making it difficult to discern details.
- Spatial Quality:** The image seems to lack a strong sense of three-dimensionality. The dragon's form and details are somewhat flat, which can make the image less immersive and realistic.
- Realism:** The image is not entirely clear, which can make it challenging to create a sense of realism. The lack of detail and depth can make the dragon appear less lifelike.
- Overall Quality:** Given the overexposure, lack of three-dimensional spatial quality, and lack of clarity, the overall quality of the image is likely to be considered low.

In conclusion, while the image has a certain artistic or stylistic quality, it does not meet the standards of high-quality CG images in terms of lighting, spatial depth, and realism. The image might be suitable for certain artistic or conceptual purposes but would not be considered high-quality in a general sense.

example: On
PASS (avg=0.903)
sim_clip=0.865, sim_reiqa=0.941

▼ View the retrieved candidate images (sorted by average similarity)

Thumbnails	Name	CLIP Sim	REIQA Sim	Average
	00390.png	0.865	0.941	0.903
	02593.jpg	0.848	0.684	0.766
	03283.jpg	0.841	0.541	0.691

How about the quality of this CG image? is it has a good quality?




Figure 4. The actual measurement of question-answering dialogues and retrieval effects on the system; the VLM is Qwen2.5-VL-7B.

Original

Q: What about the quality of this image?

A: This image is a screenshot from the video game Final Fantasy XIV. It is of a female character dressed in a red and gold outfit. The image is of poor quality, with a low resolution and a blurry background. The character's face is not clearly visible, and the details of her outfit are not well-defined. The image also has a yellowish tint, which may be due to the age of the screenshot or the quality of the original image. Overall, the image is not of high quality and may not be suitable for use in a professional or artistic context.

R4-CGQA

Q: What about the quality of this image?




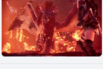
A: This image is of average quality. The character's body and hair have some flaws, but the overall painting style is not coordinated. The hair density looks thinner, which does not show the hairy texture well. In terms of material structure, the rendering effect is good, the texture of the leather bag is clear, and the shape of the cat is reasonable. However, when it comes to the detail of light and shade and the control of light intensity, this image is not quite right. The lack of delicate sense of light transition and unreasonable distribution of light intensity lead to the imbalance of light and dark contrast in some areas, which affects the overall quality and visual expression of the picture to a certain extent.

example: On

PASS (avg=0.873)

sim_clip=0.788, sim_reiqa=0.957

▼ View the retrieved candidate images (sorted by average similarity)

Thumbnails	Name	CLIP Sim	REIQA Sim	Average
	00863.jpg	0.788	0.957	0.873
	03322.jpg	0.760	0.926	0.843
	02934.jpg	0.763	0.885	0.824
	01783.jpg	0.763	0.860	0.812
	02851.jpg	0.757	0.841	0.799

What about the quality of this image?



What about the quality of this image?



Figure 5. The actual measurement of question-answering dialogues and retrieval effects on the system; the VLM is Llama 3.2-vision-11B.