

SUPPLEMENTARY MATERIALS FOR ITERATIVE SELF-IMPROVEMENT OF VISION LANGUAGE MODELS FOR IMAGE SCORING AND SELF-EXPLANATION

A. MORE DETAILS ON EXPERIMENTAL SETUP

A.1. Training and inference

Hyperparameters used in our training are listed in Table 1. The LoRA α is set to twice the LoRA rank. These parameters are selected based on validation results from the following candidates using a grid search: $\text{lr} \in \{1.e-5, 3.e-5, 5.e-5, 1.e-4, 3.e-4\}$, $\beta \in \{0.05, 0.1, 0.2\}$, $\text{rank} \in \{16, 64, 256\}$.

Images are resized to a maximum side length of 448 pixels while maintaining their original aspect ratio before being input to the model. Subsequent preprocessing follows each model’s default settings: LLaVA-1.5-7B and LLaVA-interleave-0.5B resize images to a fixed size, while LLaVA-NeXT-7B and InternVL2-2B divide them into multiple tiles.

In DPO training, we use batch size 128 and Adam optimizer for all models and datasets. The learning rate is decayed by a factor of 0.8 per iteration. In each iteration, the models are trained for one epoch. However, for the consistency training on the AVA dataset, we use a smaller subset of the data (30% of the full dataset), because the loss for the consistency training tends to decrease faster than for the score training, and the AVA dataset is quite large. All training was conducted using two A6000 GPUs (48GB memory each).

For the TIES-Merging method, we employ the one implemented in Hugging Face `peft` library with parameters: `weights=(1, 1)`, `density=0.5`, `majority_sign_method="frequency"`.

In inference and DPO data generation, texts are generated with greedy sampling and the number of max new tokens is set to 256.

For the LLM evaluation, the AVA test data is down-sampled to 1,000 instances.

A.2. Prompts

The prompt used for the DPO training and inference of the VLMs is as follows.

Please evaluate the aesthetic quality of the given photo image. The aesthetic quality should be represented by an integer score ranging from 0 to 9, with 9 being the highest score and 4 to 5 indicating a mediocre image. First, output the tag #Score followed by the aesthetic score.

Table 1. Hyperparameters used in our training.

dataset	model	lr	DPO β	LoRA rank
AVA	LLaVA-NeXT-7B	5.e-5	0.1	64
	LLaVA-1.5-7B	3.e-5	0.2	64
	InternVL2-2B	5.e-5	0.2	64
	LLaVA-interleave-0.5B	5.e-5	0.2	16
AADB	LLaVA-NeXT-7B	5.e-5	0.1	64
	LLaVA-1.5-7B	5.e-5	0.2	16
	InternVL2-2B	5.e-5	0.2	64
	LLaVA-interleave-0.5B	1.e-4	0.2	16

After that, output the tag #Explain followed by a brief explanation of the image from an aesthetic perspective. The explanation should provide the basis for the score.

The output format should be as follows:

#Score: integer

#Explain: Explanation justifying the score

After #Score:, output only the integer score with a single space before it. Do not include any additional text or symbols.

The prompt for the LLM-as-a-judge is as follows. Although we show only the consistency score in the main results, we ask the judge LLM to also evaluate the usefulness and general writing quality of explanation texts.

Your task is to evaluate the quality of explanatory texts regarding the aesthetic value of images. You will be given an aesthetic score for an image and an explanatory text justifying the score. Please rate the quality of the explanatory text on a 5-point scale (excellent, good, fair, poor, bad). The aesthetic score ranges from 0 to 9, with the following meanings: 0: Very bad 4-5: Average 9: Excellent

Evaluate the quality of the explanatory text based on the following criteria, each on a 5-point scale:

**** Consistency ****

Does the content of the explanatory text align with the aesthetic score? Is the explanation convincing as a justification for the score? Only rate as "excellent" if the alignment is perfect and the justification is highly convincing.

**** Usefulness ****

Is the content of the explanatory text useful for understanding the good points and areas for improvement of the image? Only rate as "excellent" if the text provides clear, actionable insights.

**** General ****

The overall quality of the text as a piece of writing. Only rate as "excellent" if the text is exceptionally well-written with no grammatical errors and flows logically.

Be strict in your evaluations, and only rate as "excellent" if there is no room for improvement. Return the output in JSON format only.

Example output: {"consistency": "poor", "usefulness": "good", "general": "fair"}

Below are the aesthetic score and explanatory text.

#Aesthetic score: {score}

#Explanatory text: {text}

A.3. Models

All the base model weights are taken from Hugging Face model hub:

LLaVA-NeXT-7B: <https://huggingface.co/llava-hf/llava-v1.6-vicuna-7b-hf>

LLaVA-1.5-7B: <https://huggingface.co/llava-hf/llava-1.5-7b-hf>

InternVL2-2B: <https://huggingface.co/OpenGVLab/InternVL2-2B>

LLaVA-interleave-0.5B: <https://huggingface.co/llava-hf/llava-interleave-qwen-0.5b-hf>

B. ADDITIONAL RESULTS

B.1. Complete evaluation results

In Tables 2-5, complete evaluation results for all models are shown. As LLM evaluation metrics, the usefulness metrics (*Use*) and the general writing quality (*Gen*) are presented in addition to the consistency metrics (*Cons*).

In Table 3, the LLaVA-1.5-7B models trained on consistency data occasionally show notably low consistency scores. This is primarily due to the model's outputs deviating from the required format, resulting in parsing failures and empty explanations being fed into GPT-4o. Interestingly, this issue is resolved after model merging.

B.2. More example outputs

Figures 1-4 present additional examples for model outputs sampled from the AADB test dataset for each model.

B.3. Multi-turn dialogue

When considering VLMs as AI assistants, the ability to engage in multi-turn dialogues is desirable, in addition to providing image scores and justifications. Since our fine-tuning process uses a single prompt to elicit scores and justifications, the model's multi-turn dialogue capabilities might be compromised. To assess these capabilities, we prompt the model with

"Please provide specific suggestions for improving the photo to increase its score." after it outputs the initial score and justification. This experiment uses the LLaVA-NeXT-7B model trained on the AADB dataset. An example output is shown in Fig. 5. The result indicates that the model can provide relevant responses even to questions not present in the fine-tuning data.

To quantitatively evaluate the results, we use GPT-4o to judge the usefulness, concreteness, and relevance of the generated responses with the following prompt:

Please evaluate the AI assistant's response regarding the input image. The AI assistant received the instruction "Please suggest specific methods to improve the aesthetic quality of the image" along with the image. Evaluate the response based on the following criteria:

usefulness: Whether the response is genuinely helpful in significantly improving the aesthetic quality of the image. For example, does it provide actionable advice that can lead to noticeable improvements?

concreteness: Whether the suggestions are highly specific and actionable. For example, does the response include detailed steps or techniques that can be directly applied?

relevance: Whether the suggestions are directly related to the content and context of the image. For example, are the suggestions tailored to the specific elements and style of the image?

The evaluation should be in five levels: bad, poor, fair, good, excellent. Output only in JSON format.

Example output: {"usefulness": "poor", "concreteness": "good", "relevance": "fair"}

The AI assistant's response is as follows:

{text}

The images are also provided as input during the evaluation. Table 6 presents the average scores across 500 samples from the test set. The results demonstrate that our fine-tuned model maintains performance comparable to the base model on these metrics, indicating that the fine-tuning process does not degrade its dialogue capabilities.

Table 2. Complete evaluation results of the LLaVA-NeXT-7B

(a) AVA dataset

	PLCC	SRCC	RMSE	Cons	Use	Gen	
zero-shot	0.452	0.446	0.645	3.36	2.98	3.25	
ite-1	score	0.719	0.714	0.502	3.16	2.87	3.15
	score	0.722	0.729	0.500	2.94	2.68	3.12
	consistency	0.688	0.689	0.515	3.41	2.93	3.25
ite-2	merged	0.723	0.727	0.499	3.43	2.91	3.22
	score	0.738	0.740	0.488	3.33	2.87	3.18
	consistency	0.676	0.713	0.522	3.64	2.89	3.18
ite-3	merged	0.722	0.735	0.500	3.55	2.96	3.21
	score	0.741	0.745	0.486	3.50	2.97	3.18
	consistency	0.618	0.710	0.553	3.64	2.83	3.12
ite-4	merged	0.716	0.739	0.504	3.57	2.92	3.16

(b) AADB dataset

	PLCC	SRCC	RMSE	Cons	Use	Gen	
zero-shot	0.504	0.509	0.161	3.47	2.99	3.18	
ite-1	score	0.641	0.623	0.143	3.54	3.00	3.22
	score	0.645	0.639	0.142	3.34	2.92	3.25
	consistency	0.639	0.625	0.143	3.56	3.02	3.31
ite-2	merged	0.613	0.638	0.147	3.58	3.09	3.45
	score	0.652	0.651	0.141	3.49	3.02	3.26
	consistency	0.560	0.609	0.154	3.59	3.08	3.49
ite-3	merged	0.623	0.641	0.146	3.54	3.05	3.38
	score	0.652	0.652	0.141	3.34	2.94	3.25
	consistency	0.592	0.633	0.150	3.58	3.06	3.43
ite-4	merged	0.636	0.645	0.144	3.48	2.99	3.34



Ground truth score: 9

Zero shot consistency: fair (2)

#Score: 6
#Explain: The image presents a visually appealing composition with a clear focus on the artwork. The use of color and pattern is vibrant and engaging, with a sense of depth and dimension created by the layered elements. The contrast between the brightly colored artwork and the more muted background elements draws the viewer's attention. The perspective is well-executed, leading the eye through the image and inviting further exploration. The overall aesthetic quality is high, with a balance of visual interest and harmony in the design.

Iteration-4 consistency: excellent (4)

#Score: 9
#Explain: The image captures a visually striking and harmonious composition. The interplay of geometric shapes, vibrant colors, and the playful arrangement of elements creates a sense of balance and rhythm. The use of light and shadow adds depth and contrast, enhancing the overall aesthetic quality. The image is a masterful example of modern art, demonstrating a high level of creativity and technical skill.



Ground truth score: 7

Zero shot consistency: excellent (4)

#Score: 5
#Explain: The image captures a train in motion on a track, which is a common and somewhat mundane subject. The composition is straightforward, with the train in the center and the tracks leading the viewer's eye towards the horizon. The lighting is natural and even, with no harsh shadows, which is good for the subject. However, the image lacks a strong sense of depth or perspective, and the background is not particularly interesting or engaging. The color balance is good, and the focus is sharp on the train, but there's no artistic flair or creative composition that would elevate the image beyond a simple, everyday photograph.

Iteration-4 consistency: good (3)

#Score: 7
#Explain: The image captures a powerful and dynamic scene of a freight train in motion. The composition of the photo, with the train leading the viewer's eye towards the horizon, creates a sense of movement and energy. The use of natural light and the clear blue sky in the background enhances the colors and textures of the scene, giving it a vibrant and lively feel. The focus on the train and the surrounding environment is well-balanced, and the image has a strong sense of depth and dimension. The aesthetic quality of the image is high, with a strong visual impact that conveys the beauty and power of industrial machinery in harmony with nature.



Ground truth score: 3

Zero shot consistency: good (3)

#Score: 6
#Explain: The image presents a striking contrast between the rugged, natural landscape and the man-made structures. The red rock formations, possibly a mesa, dominate the scene, their jagged edges and reddish hues adding a sense of raw, untouched beauty. The presence of greenery in the foreground provides a nice contrast to the reds, creating a harmonious balance between the natural and the man-made. The sky is clear, suggesting a calm, serene day, which enhances the overall tranquility of the scene. The composition of the image, with the rock formations in the background and the greenery in the foreground, leads the viewer's eye naturally through the scene, creating a sense of depth and perspective. The image captures the essence of the location, conveying a sense of peace and connection with nature.

Iteration-4 consistency: excellent (4)

#Score: 9
#Explain: The image captures a breathtaking view of a majestic red rock formation under a clear blue sky with a few scattered clouds. The vibrant colors of the rocks contrast beautifully with the blue of the sky, creating a visually stunning scene. The composition of the image, with the rock formation centrally placed and the vast expanse of sky surrounding it, creates a sense of awe and grandeur. The clarity and sharpness of the image are also high, allowing for fine details to be seen in the rock formation and the surrounding landscape. The image is a perfect example of natural beauty captured in a photograph.



Ground truth score: 1

Zero shot consistency: good (3)

#Score: 5
#Explain: The image has a certain level of detail and color, but the composition is somewhat cluttered with text and the focus is not sharp. The aesthetic quality is mediocre.

Iteration-4 consistency: good (3)

#Score: 5
#Explain: The image presents a visually appealing artifact with a clear focus on the subject, showcasing the intricate details and craftsmanship of the item. The use of lighting highlights the textures and colors, creating a sense of depth and dimension. The composition of the image is balanced, with the object centrally placed, drawing the viewer's attention directly to it. The background is minimal, ensuring that the viewer's attention remains on the artifact. The image effectively captures the essence of the object, demonstrating its cultural and artistic significance.

Fig. 1. Output examples of the LLaVA-NeXT-7B model on the AADB dataset.

Table 3. Complete evaluation results of the LLaVA-1.5-7B

		(a) AVA dataset			Cons	Use	Gen
		PLCC	SRCC	RMSE			
zero-shot		0.289	0.291	0.693	3.20	2.66	3.01
ite-1	score	0.702	0.694	0.516	3.15	2.59	2.93
	consistency	0.721	0.716	0.501	3.20	2.62	2.97
ite-2	merged	0.658	0.652	0.544	1.51	1.14	1.22
	score	0.700	0.694	0.517	3.35	2.77	3.04
ite-3	merged	0.723	0.718	0.500	3.30	2.69	3.00
	consistency	0.692	0.694	0.522	2.37	1.83	1.95
ite-4	merged	0.710	0.708	0.509	3.47	2.83	3.07
	score	0.730	0.723	0.495	3.23	2.67	2.98
ite-4	consistency	0.702	0.705	0.515	2.99	2.32	2.45
	merged	0.717	0.715	0.504	3.47	2.83	3.03

		(b) AADB dataset			Cons	Use	Gen
		PLCC	SRCC	RMSE			
zero-shot		0.381	0.401	0.172	3.05	2.53	2.96
ite-1	score	0.514	0.516	0.160	3.05	2.52	2.93
	consistency	0.584	0.581	0.151	2.81	2.34	2.65
ite-2	merged	0.495	0.507	0.162	0.94	0.66	0.74
	score	0.538	0.550	0.157	3.32	2.76	3.03
ite-3	merged	0.590	0.589	0.150	3.05	2.60	2.93
	consistency	0.522	0.547	0.159	3.26	2.61	2.81
ite-4	merged	0.560	0.569	0.154	3.36	2.75	3.04
	score	0.590	0.582	0.150	3.13	2.60	2.94
ite-4	consistency	0.545	0.569	0.156	3.18	2.55	2.74
	merged	0.576	0.580	0.152	3.39	2.73	3.03



Ground truth score: 9

Zero shot consistency: fair (2)

#Score: 6
#Explain: The image features a large metal sculpture of a bull with many colorful balls attached to it. The sculpture is located in a park, and the balls are arranged in various patterns and positions. The overall composition of the sculpture is visually appealing, and the use of colorful balls adds a vibrant touch to the scene. The aesthetic quality of the image is high, with a score of 6 out of 9.

Iteration-4 consistency: good (3)

#Score: 8
#Explain: The image is visually appealing with a unique and creative design. The colorful balls on the metal structure create an interesting and eye-catching display. The arrangement of the balls and the overall composition of the sculpture make it an aesthetically pleasing piece.



Ground truth score: 7

Zero shot consistency: fair (2)

#Score: 6
#Explain: The image features a train traveling down the tracks, with a large black and white engine and a few cars behind it. The train is surrounded by a beautiful blue sky, which adds a sense of depth and tranquility to the scene. The composition of the image is well-balanced, with the train taking up a significant portion of the frame while the sky serves as a captivating backdrop. The overall aesthetic quality of the image is high, with a score of 6 out of 9.

Iteration-4 consistency: good (3)

#Score: 8
#Explain: The image features a large black and silver train traveling down the tracks, with a beautiful blue sky in the background. The train is the main focus of the scene, and its size and color contrast well with the sky. The composition of the image is well-balanced, and the colors are vibrant and appealing. The aesthetic quality of the photo is high, making it an excellent representation of the subject matter.



Ground truth score: 3

Zero shot consistency: good (3)

#Score: 8
#Explain: The image captures a beautiful mountain landscape with a red rock formation. The mountain range is covered in lush greenery, and the red rock formation stands out against the vibrant greenery. The composition of the image is visually appealing, and the colors are rich and vivid. The aesthetic quality of the image is high, making it an excellent representation of the natural beauty of the landscape.

Iteration-4 consistency: excellent (4)

#Score: 9
#Explain: The image is a breathtaking view of a mountain range with red rock formations. The vibrant colors and dramatic landscape create a visually stunning scene. The composition of the image, with the mountain range stretching across the frame, draws the viewer's attention and invites them to explore the beauty of nature. The aesthetic quality of the photo is high, making it an excellent representation of the subject matter.



Ground truth score: 1

Zero shot consistency: excellent (4)

#Score: 4
#Explain: The image is a close-up of a decorative mask, possibly of a bull or a cow, with a gold and red color scheme. The mask is displayed in a museum setting, possibly in a case. The aesthetic quality of the image is relatively low due to the lack of depth and the limited focus on the mask itself. The background and the setting are not visually interesting, and the mask appears to be the main subject of the image. The overall composition and presentation of the image are not engaging or visually appealing, resulting in a score of 4 out of 9.

Iteration-4 consistency: good (3)

#Score: 8
#Explain: The image is visually appealing with vibrant colors and intricate details. The cow head mask is ornately decorated with gold accents, making it an eye-catching centerpiece. The display case adds an element of sophistication, enhancing the overall aesthetic quality of the image. The composition and lighting also contribute to the high score, creating a captivating and artistic display.

Fig. 2. Output examples of the LLaVA-1.5-7B model on the AADB dataset.

Table 4. Complete evaluation results of the InternVL2-2B

(a) AVA dataset							(b) AADB dataset						
	PLCC	SRCC	RMSE	Cons	Use	Gen		PLCC	SRCC	RMSE	Cons	Use	Gen
zero-shot	0.321	0.315	0.685	3.28	2.88	3.15	zero-shot	0.343	0.408	0.176	3.23	2.87	3.12
ite-1 score	0.677	0.675	0.532	2.87	2.56	3.04	ite-1 score	0.638	0.641	0.143	3.19	2.73	3.07
score	0.689	0.691	0.525	2.95	2.56	3.04	score	0.650	0.650	0.142	3.29	2.78	3.07
consistency	0.653	0.658	0.548	3.53	2.87	3.17	consistency	0.495	0.616	0.162	3.82	3.10	3.41
ite-2 merged	0.677	0.682	0.532	3.39	2.82	3.15	ite-2 merged	0.589	0.642	0.151	3.65	3.02	3.29
score	0.698	0.698	0.518	2.91	2.57	3.10	score	0.652	0.658	0.141	3.41	2.85	3.17
consistency	0.658	0.674	0.545	3.57	2.95	3.31	consistency	0.529	0.624	0.158	3.72	3.19	3.59
ite-3 merged	0.683	0.689	0.529	3.29	2.79	3.18	ite-3 merged	0.599	0.649	0.149	3.62	3.05	3.36
score	0.701	0.700	0.516	2.91	2.55	3.08	score	0.651	0.659	0.141	3.39	2.82	3.16
consistency	0.664	0.686	0.541	3.56	3.00	3.33	consistency	0.504	0.632	0.161	3.74	3.04	3.49
ite-4 merged	0.686	0.693	0.527	3.37	2.83	3.23	ite-4 merged	0.596	0.643	0.150	3.57	2.97	3.33



Ground truth score: 9

Zero shot consistency: good (3)

#Score: 8
#Explain: The image features a striking, modern sculpture with a vertical arrangement of colorful, cylindrical elements. The use of bright, primary colors and the geometric, repetitive pattern create a visually engaging and dynamic composition. The structure's clean lines and the contrast between the metallic elements and the colorful ones add to the aesthetic appeal. The overall effect is both bold and harmonious, making it a visually appealing piece of art.

Iteration-4 consistency: excellent (4)

#Score: 9
#Explain: The image showcases a stunningly intricate and visually captivating installation art piece. The vertical columns, adorned with colorful, geometrically patterned elements, create a mesmerizing rhythm and symmetry. The vibrant colors and the interplay of light and shadow add depth and dimension, making the piece appear almost three-dimensional. The overall composition is balanced and harmonious, evoking a sense of wonder and awe. The juxtaposition of the modern, industrial building in the background with the whimsical, colorful installation creates a striking contrast, enhancing the aesthetic appeal of the image.



Ground truth score: 7

Zero shot consistency: good (3)

#Score: 8
#Explain: The image depicts a large, black locomotive with a sleek and modern design, set against a clear blue sky. The train is positioned on a well-maintained railway track surrounded by gravel and greenery. The overall aesthetic is clean, with a focus on the train's design and the natural environment. The lighting is bright and natural, enhancing the visual appeal. The combination of the train's modern appearance and the serene background creates a visually pleasing scene.

Iteration-4 consistency: excellent (4)

#Score: 9
#Explain: The image showcases a stunningly detailed and well-lit steam locomotive, with its sleek black and white color scheme and intricate design elements. The clear blue sky and the well-maintained railway tracks enhance the overall aesthetic appeal. The contrast between the dark locomotive and the bright background creates a visually striking composition. The image exudes a sense of power and elegance, making it a visually captivating scene.



Ground truth score: 3

Zero shot consistency: excellent (4)

#Score: 8
#Explain: The image features a striking red rock formation under a clear blue sky with scattered clouds. The contrast between the vibrant red rock and the blue sky creates a visually appealing and dynamic composition. The natural beauty of the rock formation, combined with the serene sky, evokes a sense of tranquility and awe. The overall aesthetic quality of the image is high, making it a visually pleasing and memorable scene.

Iteration-4 consistency: excellent (4)

#Score: 9
#Explain: The image showcases a stunningly majestic red rock formation under a clear blue sky, with the natural beauty of the landscape creating a visually captivating scene. The contrast between the vibrant red rock and the green vegetation, along with the dramatic sky, enhances the overall aesthetic quality, making it a truly exceptional and visually stunning image.



Ground truth score: 1

Zero shot consistency: good (3)

#Score: 7
#Explain: The image features a highly detailed and ornate mask with intricate gold and red decorations. The design is elaborate and showcases a high level of craftsmanship, which contributes to a visually appealing aesthetic. The mask's design is both intricate and symmetrical, with a balance of bold colors and detailed patterns. The overall effect is one of grandeur and artistic excellence, which is a strong indicator of a high aesthetic quality.

Iteration-4 consistency: excellent (4)

#Score: 9
#Explain: The image showcases a highly detailed and intricately designed mask with vibrant colors and elaborate patterns, creating a visually striking and captivating aesthetic. The attention to detail in the facial features, the use of gold accents, and the overall composition of the mask contribute to a sense of grandeur and artistic excellence. The mask's design is both symbolic and visually appealing, making it a masterpiece of cultural artistry.

Fig. 3. Output examples of the InternVL2-2B model on the AADB dataset.

Table 5. Complete evaluation results of the LLaVA-interleave-0.5B

(a) AVA dataset

	PLCC	SRCC	RMSE	Cons	Use	Gen	
zero-shot	0.372	0.366	0.672	2.43	2.05	2.47	
ite-1	score	0.673	0.672	0.535	2.33	1.99	2.46
	score	0.697	0.695	0.519	2.21	1.93	2.51
	consistency	0.642	0.653	0.555	2.54	2.14	2.55
ite-2	merged	0.672	0.677	0.536	2.50	2.12	2.63
	score	0.697	0.697	0.519	2.24	1.92	2.43
	consistency	0.605	0.636	0.576	2.77	2.30	2.55
ite-3	merged	0.669	0.679	0.538	2.57	2.14	2.54
	score	0.698	0.699	0.518	2.30	2.01	2.48
	consistency	0.625	0.652	0.565	2.76	2.34	2.64
ite-4	merged	0.674	0.685	0.535	2.72	2.20	2.61

(b) AADB dataset

	PLCC	SRCC	RMSE	Cons	Use	Gen	
zero-shot	0.242	0.218	0.181	2.33	1.97	2.40	
ite-1	score	0.567	0.568	0.153	1.36	1.45	2.05
	score	0.597	0.602	0.149	1.57	1.50	2.20
	consistency	0.539	0.558	0.157	2.60	2.12	2.61
ite-2	merged	0.573	0.587	0.153	2.00	1.80	2.51
	score	0.588	0.607	0.151	1.66	1.56	2.26
	consistency	0.557	0.575	0.155	2.91	2.18	2.59
ite-3	merged	0.582	0.592	0.151	2.18	1.93	2.58
	score	0.607	0.612	0.148	1.41	1.37	2.08
	consistency	0.552	0.564	0.155	2.25	1.59	1.95
ite-4	merged	0.578	0.590	0.152	2.16	1.92	2.52



Ground truth score: 9

Zero shot consistency: fair (2)

#Score: 9
#Explain: The image features a sculpture of a dog in a human-like pose, with a series of cylindrical rods that resemble human legs. The sculpture is placed in front of a red building with a blue roof. The background consists of a green lawn and a few trees. The lighting is bright, and the overall color palette is warm, with the red building standing out against the greenery. The sculpture is positioned in a way that suggests a sense of movement or energy, which is a common aesthetic choice in sculptures. The image is clear and well-lit, with no visible noise or distortion.

Iteration-4 consistency: poor (1)

#Score: 4
#Explain: The image features a large, cylindrical sculpture of a dog with a rainbow-colored body. The sculpture is suspended from multiple vertical bars, creating a dynamic and visually interesting composition. The lighting is bright and even, enhancing the overall aesthetic quality of the image. The composition is balanced, with the bars framing the sculpture and the bars themselves framing the dog sculpture. The colors are vibrant and varied, adding to the visual appeal of the image.



Ground truth score: 7

Zero shot consistency: poor (1)

#Score: 9
#Explain: The image presents a clear view of a train moving along a railway track. The train is composed of several cars, including a locomotive and several freight cars. The train is moving at a moderate speed, indicated by the presence of smoke in the form of a plume of black smoke. The background is a clear blue sky, suggesting a sunny day. The image is well-lit, with the train and the sky being the main subjects. The composition of the image is balanced, with the train positioned in the center and the background being slightly blurred. The colors are vibrant, with the black and white of the train contrasting against the blue of the sky. The image does not contain any discernible text or other objects that could be considered as part of the aesthetic quality.

Iteration-4 consistency: good (3)

#Score: 4
#Explain: The image is a close-up of a train engine, with a clear view of the engine's design and the smoke coming from it. The background is blurred, but it appears to be a train track with some industrial structures in the distance. The lighting is bright and even, which suggests a good day for train travel. The composition is simple and focused, with the engine being the main subject and the background being less distracting. The image does not contain any other significant elements that would detract from the main subject.



Ground truth score: 3

Zero shot consistency: excellent (4)

#Score: 9
#Explain: The image is a clear and well-lit photograph of a mountainous landscape. The colors are vibrant, and the contrast between the blue sky and the red and brown hues of the rocks and vegetation is striking. The composition is well-balanced, with the mountainous terrain serving as the main focus. The image does not contain any visible noise or artifacts, and the focus is sharp, allowing for clear visibility of the details of the landscape. The lighting is even, and the exposure is appropriate, ensuring that the image is not overexposed or underexposed. The image does not contain any distracting elements, and the composition is aesthetically pleasing.

Iteration-4 consistency: poor (1)

#Score: 4
#Explain: The image is a clear and well-lit photograph of a mountainous landscape with a prominent rock formation. The colors are vibrant, and the contrast is high, making the image visually appealing. The composition is well-balanced, with the mountainous terrain serving as a backdrop and the rock formation as the focal point. The image does not contain any distracting elements or obstructions, and the focus is sharp, allowing for a clear view of the details of the landscape.



Ground truth score: 1

Zero shot consistency: fair (2)

#Score: 9
#Explain: The image is a close-up of a red and gold-brown horned headpiece with a detailed design. The horns are intricately carved, and the horns are positioned in a way that suggests they are being lifted or turned. The image also contains text that provides information about the horn, such as "Topping Level: Mark The Confetti," "Topping Level: Top Bubbles," and "Topping Level: Top Bubbles." The image is well-lit, with a clear focus on the horn, and the colors are vibrant and saturated. The composition is simple, with the horns being the main subject. The image does not contain any other objects or text that can be described.

Iteration-4 consistency: poor (1)

#Score: 3
#Explain: The image is a close-up of a red and gold-brown horned head, likely a type of traditional or cultural headgear. The horns are intricately designed with patterns and details, and the horns are positioned in a way that suggests they are being held or used. The image is clear and well-lit, allowing for a detailed view of the horns. The composition is simple, with the horns being the main focus. The image does not contain any distracting elements or objects that would detract from the main subject. The overall aesthetic quality is good, with a clear and focused subject.

Fig. 4. Output examples of the LLaVA-interleave-0.5B model on the AADB dataset.

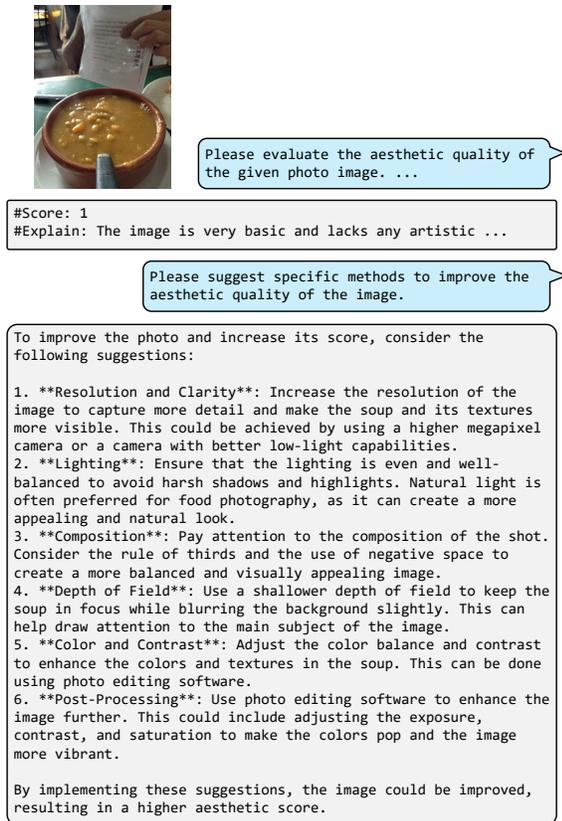


Fig. 5. Example of a multi-turn dialogue

Table 6. Evaluation of dialogue capabilities for the LLaVA-NeXT-7B model on the AADB dataset.

	usefulness	concreteness	relevance
base model	3.23	3.30	3.71
trained model	3.22	3.28	3.66