

ICM-Assistant: Instruction-tuning Multimodal Large Language Models for Rule-based Explainable Image Content Moderation

Supplementary Materials

Disclaimer: This supplementary material contains harmful content that may be disturbing to some readers.

Section A: Dataset

Attribute products examples of S-ICM

We provide the full list of attribute products in the dataset for sexy ICM and show some of the image examples. In Fig 1 shown below, 35 attribute products and example images are shown for the sexy ICM task. Four attribute products (as negative samples) are marked with green boundaries.

Other data examples

We also provide example images for other validation and training sets mentioned in the main paper. As shown in Fig. 2 of this supplementary material, (a) shows example images for the ICM-Instruct dataset on the moderation term “Horrifying”; (b) shows example images for the ICM-Instruct dataset on the moderation term “Gambling”; (c) shows example AIGC images of in the ICM-Test set generated with SDXL (Podell et al. 2023); and (d) shows example UGC advertisement images collected from the Internet.

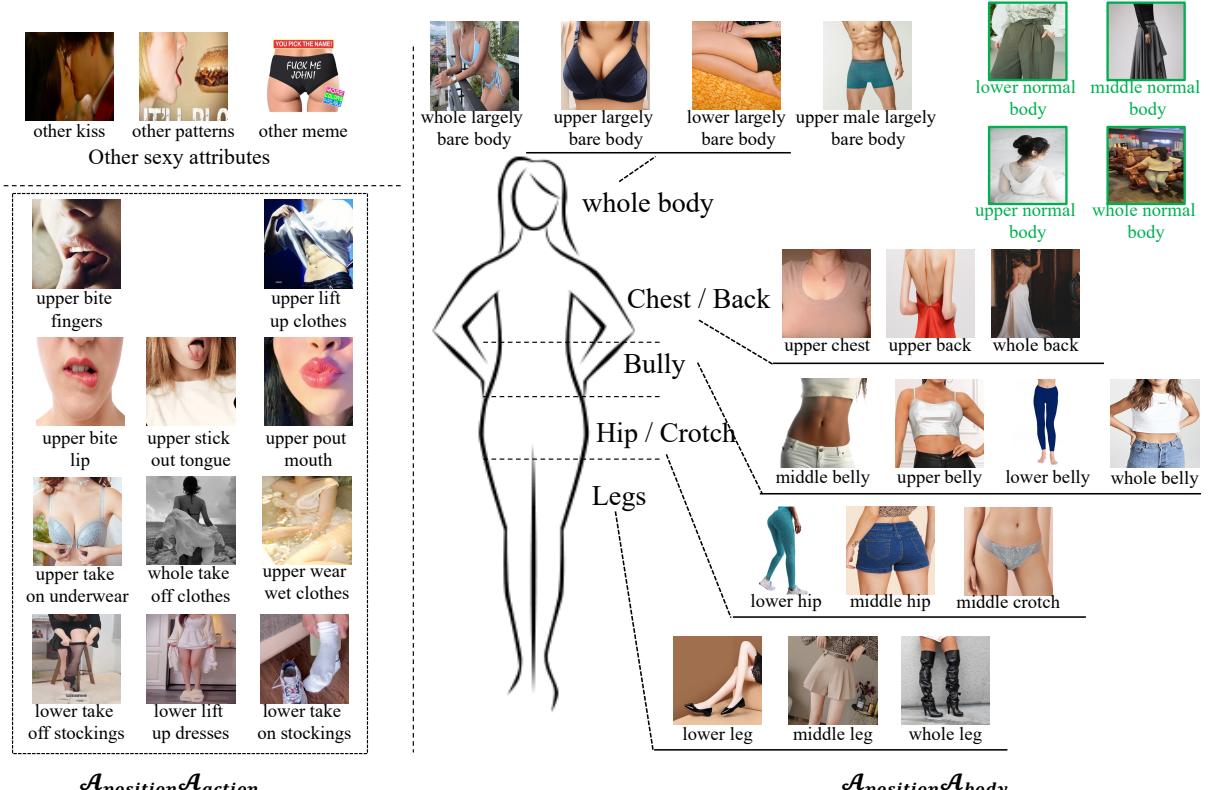


Figure 1: Attribute products and example images in ICM-Instruct for the sexy ICM task.

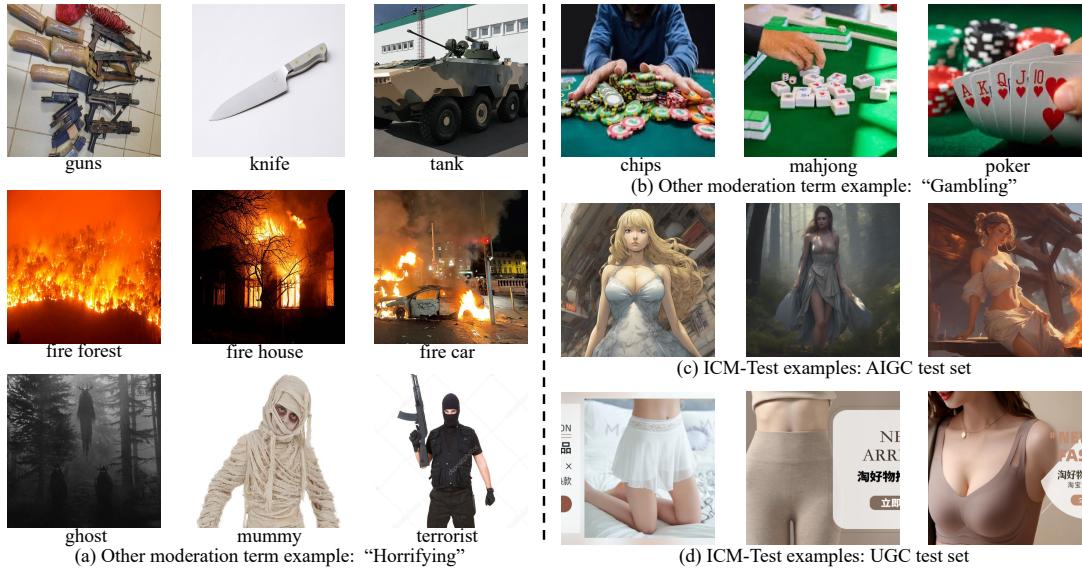


Figure 2: Other training and validation sets mentioned in the main paper: (a) example images for the ICM-Instruct dataset on the moderation term “Horrifying”; (b) example images for the ICM-Instruct dataset on the moderation term “Gambling”; (c) example AIGC images of the ICM-Test set generated with SDXL (Podell et al. 2023); and (d) example UGC advertisement images collected from the Internet.

Section B: Method details

Detailed prompts in the ICM-Instruct dataset generation pipeline

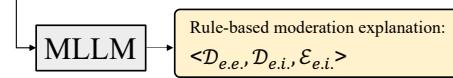
We provide (1) the description-enrich prompts for generating the moderation explanations. We sum up the three-step prompts in one prompt to accelerate the data generation process, and (2) the prompts for generating the moderation Q-A from the moderation explanations. The system message for the LLM is set as '*You are an image content moderator who can analyze a single image.*'

List of explicit descriptions

We provide explicit descriptions (examples) prepared by human moderators, for generating the moderation explanations mentioned in the main paper. In Table 1 of this supplementary material, an example of explicit description is provided, we made small adjustments according to the images in the description-enrich prompts.

<i>I</i>		Answer the following three questions sequentially based on the given description. Do not use the first person, answer with an objective description:
\mathcal{D}_e		Explicit description: It is known that the person in the image is naked from thighs to ankles, and this image is sexy. (example)
$\mathcal{P}_{e.e}$	1. Please describe the image in detail with explicit information such as person's movements/clothes.	
$\mathcal{P}_{e.i}$	2. Please describe the implicit content of the image in detail, such as overall atmosphere.	
\mathcal{P}_{exp}	3. Explain why this image is or not sexy based on the image explicit content and the implicit content.	

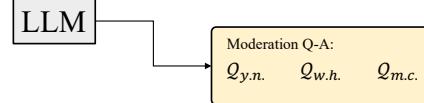
$$\mathcal{P}_{in}^1 = < \textcolor{brown}{I} \textcolor{brown}{C} \mathcal{D}_e \mathcal{P}_{e.e} \mathcal{P}_{e.i} \mathcal{P}_{exp} >$$



(a) Rule-based moderation explanation generation prompts

$\mathcal{P}_{Q.A.}$	For multiple-choice question: You receive a description of an image. Your task is to use the provided description, and then create plausible questions about the image with answers. The questions include 5 multiple-choice questions (4 choices, A, B, C, D, per question). The response is in the format of a table. The columns are 'Question', 'Choices' (only for multiple-choice questions), and 'Answer'. All answers can be derived from the description and should be concise and accurate.
	For Yes/No and What/How question: You receive a description of an image. Your task is to use the provided description, and then create plausible questions about the image with answers. The questions include 6 yes/no questions, 2 what questions, and 2 how questions. The response is in the format of a table. The columns are 'Type of Question', 'Question', and 'Answer'. All answers can be derived from the description and should be concise and accurate.

$$\mathcal{P}_{in}^2 = < \mathcal{D}_{e.e} \mathcal{D}_{e.i} \mathcal{E}_{e.i} \mathcal{P}_{Q.A.} >$$



(b) Moderation Q-A generation prompts

Figure 3: Detailed prompts in the ICM-Instruct dataset generation pipeline. (a) The description-enrich prompts for generating the moderation explanations. (b) The prompts for generating the moderation question-answering pairs from the moderation explanations.

Attribute product	Explicit description
upper_chest	It is known that the image is a close-up of a person's upper chest, showing her cleavage.
upper_back	It is known that the person in the image has her back bare, which often gives people a sexy feel.
upper_belly	It is known that the person in the image has her belly exposed.
upper_lift_up_clothes	It is known that the person in the image is lifting up clothes, which often gives people a sexy feel.
upper_stick_out_tongue	It is known that the woman in the image is sticking out her tongue, which often gives people a sexy feel.
upper_pout_mouth	It is known that the person in the image is posing as if kissing, which often gives people a sexy feel.
upper_take_on_underwear	It is known that the woman in the image is putting on the underwear, which often gives people a sexy feel.
upper_bite_lip	It is known that the person in the image is biting her lip, which often gives people a sexy feel.
upper_male_largely_bare_body	It is known that the male in the image has a large area of his body exposed.
upper_largely_bare_body	It is known that the person in the image has a large area of her body exposed.
upper_bite_fingers	It is known that if a person puts her finger inside her mouth, it often gives people a sexy feel.
upper_wear_wet_clothes	It is known that the person's upper clothes are soaked, which often gives people a sexy feel.
middle_legs	It is known that if a person's clothes almost expose her thighs, it often gives people a sexy feel.
middle_belly	It is known that the person in the image has her belly exposed.
middle_hip	It is known that the image is a close-up of the buttocks.
middle_crotch	It is known that the image is a close-up of the crotch, and the person is only wearing underwear.
lower_leg	It is known that the person in the image is naked from her thighs to his/her ankles.
lower_belly	It is known that the person in the image has her belly exposed.
lower_hip	It is known that the person in the photo deliberately shows her buttocks, an action that often gives people a sexy feel.
lower_take_on_stockings	It is known that the person in the image is taking on lower stockings, which often gives people a sexy feel.
lower_largely_bare_body	It is known that the person in the image has a large area of her body exposed.
lower_take_off_stockings	It is known that the woman in the image is taking off lower stockings, which often gives people a sexy feel.
lower_lift_up_dresses	It is known that the person in the image is lifting their clothes, which often gives people a sexy feel.
whole_largely_bare_body	It is known that the person in the image has a large area of her body exposed.
whole_leg	It is known that if a person's clothes almost expose her thighs, it often gives people a sexy feel.
whole_belly	It is known that the person in the image has her belly exposed.
whole_back	It is known that the person in the image has her back bare, which often gives people a sexy feel.
whole_take_off_clothes	It is known that the person in the image is taking off clothes, which often gives people a sexy feel.
other_meme	It is known that there are texts with teasing or sexually suggestive meanings in the image.
other_patterns	It is known that there are sexually suggestive patterns or similar objects in the image.
other_kiss	It is known that the people in the image are kissing.
upper_normal_body	It is known that the exposure level of this image is very low and it is an appropriate image.
middle_normal_body	It is known that the exposure level of this image is very low and it is an appropriate image.
lower_normal_body	It is known that the exposure level of this image is very low and it is an appropriate image.
whole_normal_body	It is known that the exposure level of this image is very low and it is an appropriate image.

Table 1: Example explicit descriptions for each attribute product.

Section C: Experiments

Prompts for Evaluation MEQ

We provide the prompt for evaluating the quality of the moderation explanation in Fig. 4 shown in this supplementary material. We use GPT-3.5 (OpenAI 2024) as our examiner with the system message as ‘You are a helpful and precise assistant for checking the quality of the answer.’ and prompt it with the instructions (See Fig. 4 blue region).

Quantitative results on larger MLLMs

We provide more quantitative results on larger scale MLLM models: LLaVA-v1.5-13B (Liu et al. 2024), LLaVA-ShareGPT4V-13B (Chen et al. 2023) and CogVLM2-19B (Wang et al. 2023).

Model	ICM acc.	MEQ	MQA acc.
LLaVA-v1.5-13B	0.894	0.725	0.809
ICM-LLaVA-v1.5-13B	0.981	0.836	0.924
ShareGPT4V-13B	0.860	0.672	0.831
ICM-ShareGPT4V-13B	0.986	0.847	0.927
CogVLM2-19B	0.896	0.717	0.822
ICM-CogVLM2-19B	0.975	0.829	0.911

Table 2: Results on ICM accuracy, MEQ, and MQA accuracy on the 13B and 19B MLLMs on the ICM-Val set.

Qualitative comparison on moderation

Q-A

We provide more examples of moderation Q-A in Fig. 6 of this supplementary material, by comparing our ICM-LLaVA-v1.5 (green) with LLaVA-v1.5 (Liu et al. 2024) (both are 7B models).

Qualitative comparison on moderation explanations

We provide more examples of moderation explanation in Fig. 5 of this supplementary material, by comparing our ICM-LLaVA-v1.5 (green) with LLaVA-v1.5 (Liu et al. 2024) (both are 7B models).

Ablations on the Scale of ICM-Instruct (explanation data)

We provide more examples on the moderation explanation results with increasing amount of moderation explanation data in the ICM-Instruct dataset in Fig. 7 of this supplementary material.

MEQ: Evaluating moderation explanations quality with LLM

Q: Is the image sexy?
Explain the reason.



ICM-Assistant Model (Rule 1)



Explanation: The image shows ... woman wearing short skirt and high heels ... This can be considered inappropriate and offensive in conservative cultures and contexts, especially when children are involved.
Conclusion: The picture is **sexy**.

Human moderator



Explanation: The woman exposes her legs, which is considered as sexy attribute for younger children, meanwhile, exposing legs is not allowed in the conservative culture.
Conclusion: The picture is **sexy**.

R1 13- 

Prompt for evaluating the quality of the moderation explanation

Task: You will receive two descriptions, both explaining why the same image is sexy or not sexy. One of the descriptions is written by a real person, indicated by the title 'Real person's description'. The other description is generated by an algorithm, indicated by the title 'Algorithm's description'.

Please read the two descriptions carefully and evaluate the quality of 'Algorithm's description' based on the content of 'Real person's description'. Note that: 1. the 'Real person's description' is accurate; 2. do not only compare the lengths of two descriptions. The response is in the format of a table, where there are two columns: "Criteria" and "Score". Under column "Criteria", there are four rows: "Accuracy", "Preciseness", "Comprehensiveness", and "Fluency". You need to fill in their scores under the column "Score", which are in 0-100. Below are definitions:

- "Accuracy" score: This measures how closely the "Algorithm's description" matches the "Real person's description" in explaining the image.
- "Preciseness" score: It measures the inclusion of details in the "Algorithm's description" that are not mentioned or implied in "Real person's description". If "Algorithm's description" includes fewer explanations that are not spelled out in "Real person's description", the "Preciseness" score is higher.
- "Comprehensiveness" score: This measures whether the "Algorithm's description" covers all aspects mentioned in "Real person's description".
- "Fluency" score: It measures how fluent of the presentation of the "Algorithm's description".



LLM

Criteria	Score
Accuracy	90
Preciseness	70
Comprehensiveness	80
Fluency	85

Figure 4: Detailed prompts for evaluating the quality of the moderation explanations (MEQ).

Q: Is the image sexy?
Explain the reason.



Rule Description

LLaVA

Explanation: The woman in the image is wearing a beige bra and is in the process of putting it on. She is holding the straps of the bra with her hands and adjusting them. The image is sexy because the woman is wearing a beige bra, which is a color often associated with sensuality and allure. Additionally, her actions of putting on the bra suggest a sense of intimacy and vulnerability, which can be perceived as sexy. The combination of the woman's attire and her actions creates an atmosphere that is alluring and enticing.

Conclusion: The picture is **sexy**.

ICM-LLaVA

Explanation: The image features a woman wearing a white bra and a pink undergarment. She is adjusting her bra, possibly fixing a strap or ensuring it fits properly. The image portrays a woman who is comfortable in her own skin and confident in her appearance. The act of adjusting her bra and smiling at the camera demonstrates her self-assurance and self-care. The image can be interpreted as a celebration of feminine beauty and the importance of self-love and self-care.

Conclusion: The picture is **not sexy**.

Rule Description

LLaVA

Explanation: The woman in the image is wearing a beige bra and is in the process of putting it on. The image is sexy because it showcases the woman's body and the act of putting on a bra, which is often associated with intimacy and sensuality. The light blue background provides a contrasting color that further accentuates the woman and her actions. The beige bra she is wearing adds a soft and delicate touch to the image, while the overall atmosphere conveys a sense of allure and seduction.

Conclusion: The picture is **sexy**.

ICM-LLaVA

Explanation: The image shows a young woman taking off their clothes in a bathroom. She is wearing a blue and white checkered bra and matching underwear. She is holding their shirt open with both hands, exposing their torso and chest. The image is sexy due to the person's actions and attire. She is taking off their clothes in a bathroom, which is a private and intimate setting. The act of removing one's clothing often conveys a sense of vulnerability and sensuality. She is wearing a blue and white checkered bra, which is a revealing and provocative choice of clothing.

Conclusion: The picture is **sexy**.

LLaVA

Explanation: The image features a woman wearing a bikini, standing in a bathroom and holding her shirt open. The scene is suggestive and provocative, as it showcases the woman's body and her attire in a private setting. The image raises questions about the objectification of women and the societal expectations of beauty. It can be interpreted as a representation of the pressure on women to conform to certain standards of appearance, which may lead to the exploitation of their bodies for commercial purposes.

Conclusion: The picture is **sexy**.

ICM-LLaVA

Explanation: The woman in the image is taking off her clothes, specifically her blue shirt and white bra. She is standing in a bathroom, with a sink and a mirror visible in the background. The image is sexy because the woman is exposing her body, which is a common element in sexy images. Additionally, the bathroom setting adds to the intimate atmosphere, as it is a private space where people typically feel comfortable and vulnerable. The combination of the woman's actions and the setting creates a sense of allure and sensuality, making the image sexy.

Conclusion: The picture is **sexy**.

Rule Description

LLaVA

Explanation: The image features a person wearing a white dress, which is soaked and clinging to their body. The overall atmosphere of the image is intimate and sensual, as the person's wet clothes and the close-up perspective create a sense of vulnerability and allure. The image is sexy because it captures a moment of intimacy and allure, with the person's wet clothes and the close-up perspective emphasizing the sensuality of the scene. The close-up perspective allows the viewer to focus on the person's body and the details of their wet clothes, creating a sense of intimacy and allure.

Conclusion: The picture is **sexy**.

ICM-LLaVA

Explanation: The image features a woman in a white dress, lying on her side in a pool of water. She is wearing a white top and appears to be posing for the camera. The scene is quite suggestive and can be considered as a sexy content. The fact that she is lying on her side in a pool of water adds to the sensuality of the scene. The woman's pose and the overall setting create an atmosphere of relaxation and seduction.

Conclusion: The picture is **not sexy**.

Rule Description

LLaVA

Explanation: The image features a woman sitting in a bathtub filled with water. The image is sexy because of the woman's attire and the intimate setting. The white dress, which is soaked, reveals her skin and adds a sense of allure to her appearance. The bathtub setting, filled with water, creates a sense of privacy and relaxation, which further enhances the sexy atmosphere. The woman's gaze off to the side adds a sense of mystery and intrigue, making the viewer curious about what she is looking at.

Conclusion: The picture is **sexy**.

ICM-LLaVA

Explanation: This image features a woman with long hair, wearing a black top, who is biting a banana. With a banana in her right hand, she bites it with her mouth open. The woman's head is tilted back and she appears to be enjoying a banana. The image is sexy because it uses a common sexual overtones to create a sense of intimacy and pleasure. The action of a woman biting into a banana, which is used as a symbol of penis, evokes the idea of oral sex. It's a taboo, often erotic act. The image creates a sense of excitement and desire, which is often associated with sexual arousal.

Conclusion: The picture is **sexy**.

LLaVA

Explanation: The image features a woman with long hair, wearing a black shirt, and she is eating a banana. The woman is making a face as she bites into the banana, which adds to the sensuality of the scene. The close-up shot of the woman's face and the banana emphasizes the focus on the interaction between the woman and the fruit. From the image, it can be concluded that the woman is enjoying the taste and texture of the banana, and her facial expression adds a sense of sensuality to the scene. The close-up shot and the woman's attire suggest that the image might be intended to evoke a sense of intimacy or allure.

Conclusion: The picture is **sexy**.

ICM-LLaVA

Explanation: The image features a woman with blonde hair, wearing a black top, who is licking a banana. The banana is positioned in a way that it appears to be a male penis. The image is sexy because it uses visual cues to create a sexual context. The woman's actions of licking the banana and the positioning of the banana as a male penis evoke strong sexual implications. The overall atmosphere of the image is sensual and provocative, which can be considered sexy based on the content and the overall atmosphere.

Conclusion: The picture is **sexy**.

Figure 5: Comparing moderation quality of ICM-LLaVA-v1.5 and its base model. Our method can provide consistent answers with those from human moderators and explain well.

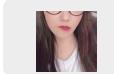
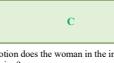
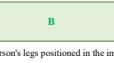
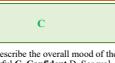
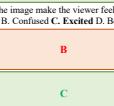
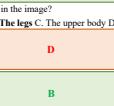
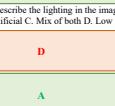
				Legend	Answer	Consistent	Inconsistent		
R1	13	What is the woman in the image doing with her lips? A. Pursing them B. Smiling widely C. Biting them D. Kissing someone.		LLaVA	D	C	C		
LLaVA-ICM		What is the person in the image lifting? A. Hat B. Shirt and pants C. Skirt and jacket D. Book		LLaVA	C	B	B		
		What is the person in the image wearing on their feet? A. High heels B. Flats C. Boots D. Socks.		LLaVA	C	B	B		
		What is the person's posture in the image? A. Formal B. Casual C. Athletic D. Elegant		LLaVA	B	C	C		
		What emotion does the woman in the image appear to be expressing? A. Happiness B. Sadness C. Anger D. Seductiveness		LLaVA	C	A	D		
		How is the person holding shirt up in the image? A. With their teeth B. With their hands C. With a stick D. With a scarf		LLaVA	B	B	C		
		How are the person's legs positioned in the image? A. Crossed B. Spread apart C. Folded D. Together		LLaVA	C	A	D		
		How would you describe the overall mood of the image? A. Serious B. Playful C. Confident D. Sensual		LLaVA	B	D	D		
		Is the image sexy? (Yes)		LLaVA	No	B	D		
		How does the image make the viewer feel? A. Anxious B. Confused C. Excited D. Bored.		LLaVA	D	D	A		
		What is visible in the image? A. The face B. The legs C. The upper body D. The feet		LLaVA	C	B	A		
		How would you describe the lighting in the image? A. Natural B. Artificial C. Mix of both D. Low light		LLaVA	D	D	D		
		Q: Is the image sexy? Explain the reason.		Explanation:	The woman in the image is laying on her stomach in a body of water. She is wearing a white shirt that is soaked and appears to be wet. The image is sexy because the woman is in a vulnerable position, laying on her stomach in a body of water. Her wet clothes create a sense of intimacy and vulnerability. The combination of these elements, along with the overall atmosphere of the image, contributes to the sexy feel of the picture.	Explanation:	The image features a woman lying on her back in a body of water. She is wearing a white blouse that is soaked, giving her a sexy appearance. Her hair is wet, and she is looking down at the water. The overall atmosphere of the image is intimate and sensual. The image is sexy because of the woman's wet clothes and the intimate atmosphere created by the water. The woman's soaked blouse and wet hair suggest that she has been in the water for some time, adding to the sensuality of the image. The fact that she is lying on her back in the water further adds to the sensual atmosphere, as it suggests a level of trust and openness.	Explanation:	The image features a woman lying on her back in a body of water. She is wearing a white blouse that is soaked, giving her a sexy appearance. Her hair is wet, and she is looking down at the water. The overall atmosphere of the image is intimate and sensual. The image is sexy because of the woman's wet clothes and the intimate atmosphere created by the water. The woman's soaked blouse and wet hair suggest that she has been in the water for some time, adding to the sensuality of the image. The fact that she is lying on her back in the water further adds to the sensual atmosphere, as it suggests a level of trust and openness.
		Q: Is the image sexy? Explain the reason.		Explanation:	The image features a woman with a finger inside her mouth. Her facial expression is serious. The image is sexy due to the woman's serious expression and the fact that she is putting her finger inside her mouth. This action is often associated with sensuality and intimacy. Additionally, the woman's serious expression adds to the overall atmosphere of the image, making it more alluring and intriguing.	Explanation:	The image features a close-up of a person's face, with their mouth open and tongue sticking out. The image is sexy because it captures a moment of intimacy and sensuality. The person's mouth is open, and their tongue is sticking out, which is often associated with a seductive or provocative gesture. The close-up nature of the image allows the viewer to focus on the person's facial features and expressions, further emphasizing the sensual nature of the scene.	Explanation:	The image features a close-up of a person's face, with their mouth open and tongue sticking out. The focus is on the person's mouth and tongue. The image is sexy because it portrays a sensual and intimate moment. The person's open mouth and tongue sticking out suggest a sense of vulnerability and sensuality. The close-up of the person's face emphasizes the focus on the mouth and tongue. The overall atmosphere of the image is intimate and provocative, which can evoke feelings of desire and arousal in the viewer.
		Q: Is the image sexy? Explain the reason.		Explanation:	The image features a person wearing white shorts and a pink top. The shorts are cropped at the waist, revealing the person's thighs. The image is sexy because it showcases the person's thighs, which are often associated with sensuality and allure.	Explanation:	The image is a close-up of a person's crotch, and the person is only wearing white shorts. The image is sexy because it focuses on the person's lower body, specifically the crotch area, which is often associated with sensuality and intimacy. The white shorts worn by the person add a contrast to the skin tone, which can be seen as a symbol of purity and innocence, but it also draws attention to the person's body.	Explanation:	The image is a close-up of a person's crotch, and the person is only wearing underwear. The focus is on the person's exposed skin and the underwear. The image is sexy because it is a close-up of a person's crotch, which is a private and intimate part of the body. The person is only wearing underwear, which is a common garment associated with sensuality and intimacy. The close-up nature of the image emphasizes the person's skin and the underwear, creating a sense of intimacy and allure.

Figure 6: Comparing moderation Q-A results of ICM-LLaVA-v1.5 (green) and its base model (red). Our method can provide correct answers on the questions related to the moderation term “sexy”.

Input image	20% Explanation	60% Explanation	100% Explanation
	<p>Explanation: The woman in the image is laying on her stomach in a body of water. She is wearing a white shirt that is soaked and appears to be wet. The image is sexy because the woman is in a vulnerable position, laying on her stomach in a body of water. Her wet clothes create a sense of intimacy and vulnerability. The combination of these elements, along with the overall atmosphere of the image, contributes to the sexy feel of the picture.</p>	<p>Explanation: The image features a woman lying on her back in a body of water. She is wearing a white blouse that is soaked, giving her a sexy appearance. Her hair is wet, and she is looking down at the water. The overall atmosphere of the image is intimate and sensual. The image is sexy because of the woman's wet clothes and the intimate atmosphere created by the water. The woman's soaked blouse and wet hair suggest that she has been in the water for some time, adding to the sensuality of the image. The fact that she is lying on her back in the water further adds to the sensual atmosphere, as it suggests a level of trust and openness.</p>	<p>Explanation: The image features a woman lying on her back in a body of water. She is wearing a white blouse that is soaked, giving her a sexy appearance. Her hair is wet, and she is looking down at the water. The overall atmosphere of the image is intimate and sensual. The image is sexy because of the woman's wet clothes and the intimate atmosphere created by the water. The woman's soaked blouse and wet hair suggest that she has been in the water for some time, adding to the sensuality of the image. The fact that she is lying on her back in the water further adds to the sensual atmosphere, as it suggests a level of trust and openness.</p>

Figure 7: Moderation explanation results from models trained with increasing amount of moderation explanation data. 20%, 60%, and 100% explanation indicate the amount of explanation data (part of the training set of the baseline models) employed in model training. The more detailed explanation provided in the training set, the more accurate, fluent, and comprehensive the explanation can be provided in the ICM tasks. (The explicit (green) and implicit (bold) explanations are marked in the images.)

References

- Chen, L.; Li, J.; Dong, X.; Zhang, P.; He, C.; Wang, J.; Zhao, F.; and Lin, D. 2023. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.
- OpenAI. 2024. ChatGPT. <https://openai.com/blog/chatgpt/>.
- Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. *arXiv:2307.01952*.
- Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; Song, X.; et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.