

Pixels Don't Lie (But Your Detector Might): Bootstrapping MLLM-as-a-Judge for Trustworthy Deepfake Detection and Reasoning Supervision

Supplementary Material

6. Ethics Statement

Our work raises important ethical considerations, particularly due to the use of visual data and human participation in deepfake research. The dataset and human studies were conducted following institutional guidelines and community standards to ensure responsible and transparent research. Below, we outline all ethical considerations related to this work.

Use of Public and Generated Visual Data: All images used in DeepfakeJudge originate from publicly available datasets or from synthetically generated sources. The real subset is derived from Open-Images V7 [18] from Google, used under its research license with appropriate attribution. The synthetic and edited subsets were created using text-to-image and image-editing models, including Gemini [9], SeedDream [36], Flux-Kontext [19], and Qwen-Edit [48], solely for academic and non-commercial research. Each generated or edited image underwent linguistic, semantic, and NSFW filtering, followed by manual verification to remove inappropriate, offensive, or personally identifiable content. This ensures that all data are ethically sourced and safe for research use. The collection and use of such material are consistent with established practices followed in peer-reviewed benchmarks such as FaceForensics++ [34] and DFDC [10], which also use public or generated content for research under fair-use provisions.

Human Annotation: The first phase of human annotation involved six trained annotators who labeled reasoning cues, bounding boxes, and referring expressions across both in-distribution and out-of-distribution subsets. All annotators were over 18 years of age, affiliated with university research groups, and completed a shared pilot phase to ensure consistency. Inter-annotator agreement reached substantial alignment (Cohen's $\kappa = 0.71$). All annotators were compensated for their time and effort. No personal information was collected, and all materials were screened to exclude NSFW or sensitive content. Participation was voluntary, and all annotations were used exclusively for research purposes. A second phase of annotation was conducted to create the DeepfakeJudge-Meta-Human dataset. Two expert annotators independently evaluated reasoning quality and pairwise reasoning preferences. Both annotators were over 18 years old, recruited through academic networks, and compensated for their work. We evaluated overlapping

subsets to ensure reliability, resulting in high inter-annotator agreement (Cohen's $\kappa = 0.80$ for pairwise and $MSE = 0.39$ for pointwise evaluation). All participation was voluntary, and no personal data were collected.

User Study: The user study evaluating reasoning preferences followed institutional IRB guidelines (IRB Proposal No: MBZUAI/IRB/2025/0012). Ten adult participants were recruited from research groups at the authors' affiliated universities. Participation was voluntary, and participants were compensated. Before participation, each subject received an explanatory statement outlining the study objectives and procedures, along with a few examples detailing the procedure, and informed consent was obtained through the study form. All materials used in the study were screened to exclude NSFW or distressing content. No personal information or identifying data were recorded, and all responses were anonymized.

Access and Use Policy: All datasets, models, and code will be released for academic, non-commercial use only, under a strict End-User License Agreement (EULA). Access will be granted only to verified academic researchers under the following conditions: (1) Use is limited to academic, educational, and not-for-profit research. (2) Each institution accepts responsibility for its authorized users. (3) Redistribution, modification, or misuse of the data is prohibited. (4) Access may be revoked or modified by licensors at any time. (5) Any use that could cause embarrassment, harm, or distress to subjects is strictly forbidden.

This controlled-access policy aligns with established community practices, such as Faceforensics++ [34] and DFDC [10], ensuring that research on deepfake detection and reasoning remains ethical, traceable, and responsibly conducted. All stages of this work, including data creation, annotation, and user evaluation, adhered to institutional and community standards for human research ethics.

7. Effect of Paraphrasing

Tables 5, 6 show the relative performance of DeepfakeJudge-3B model, trained with and without paraphrasing. These results signify the importance of paraphrasing to avoid overfitting on linguistic styles, thereby forcing the model to focus on semantic information instead.

Table 5. Performance of DeepfakeJudge-3B on the DeepfakeJudge-Meta dataset with and without paraphrasing.

Model	Para	RMSE (\downarrow)	MSE (\downarrow)	s (\uparrow)	p (\uparrow)
DeepfakeJudge-3B	✓	0.69	0.48	0.92	0.92
DeepfakeJudge-3B	✗	1.39	1.93	0.60	0.62

Table 6. Pairwise accuracy of DeepfakeJudge-3B on the DeepfakeJudge-Meta dataset with and without paraphrasing.

Model	Paraphrasing	DFJM (%)
DeepfakeJudge-3B	✓	94.4
DeepfakeJudge-3B	✗	89.1

8. Robustness to Common Image Degradations.

Table 7. Robustness under common image degradations.

Model Variant	Pointwise		Pairwise
	RMSE \downarrow	Pearson \uparrow	Accuracy (%) \uparrow
JPEG-70	0.64	0.93	95.0
JPEG-80	0.65	0.93	95.1
Resize-0.5	0.65	0.92	95.1
Resize-0.75	0.62	0.93	95.4
DeepfakeJudge-7B	0.61	0.93	96.2

We conduct additional experiments applying JPEG compression (quality levels: 70 and 80) and resizing (scale factor: 0.5 and 0.75). The results are summarized in Table 7. The model’s stability against common social media perturbations highlights its robustness and practical applicability.

9. Model Bias and Closed-Loop Confirmation Bias

Our results show that the judge does not mimic GPT’s preferences or style. Although GPT is used to verbalize gold rationales from dense human annotations, evaluation is conducted on independently human-annotated benchmarks, where DFJ consistently outperforms GPT in correlation, calibration, and pairwise agreement (DFJM-Human subset, Tables 3 and 4), which would not occur if it simply reproduced GPT’s evaluation behavior.

We further validate generalization via an out-of-loop evaluator test: We re-score degraded responses with Gemini-2.5-Flash as the evaluator, which yields high agreement with target ratings (89% accuracy, Pearson 0.96), comparable to GPT as evaluator (92%, Pearson 0.97), which shows that the generated degraded outputs are grounded in image rather than model bias. Moreover, GPT does not receive abnormally high reasoning scores from

DFJ (Table 2, row 2), consistent with its weak detection performance (Table 1, row 2). Finally, style imitation is mitigated by paraphrasing accepted rationales into multiple surface forms (Figure 5), showing that DFJ learns a human-aligned notion of visually grounded reasoning rather than GPT-specific preferences.

Further, our pipeline does not suffer from closed-loop confirmation bias. The generator and evaluator are explicitly decoupled: the evaluator never sees the generator’s intended quality level, and the generator never receives the evaluator’s numeric score; acceptance is enforced externally, with refinement guided only by textual critiques. Empirically, we test for circularity using an out-of-loop evaluator by re-scoring responses with Gemini. Although refinement is performed by GPT, Gemini’s agreement with the target ratings increases from 89% to 95%, showing that refinement improves rating accuracy in a way that generalizes beyond the in-loop evaluator’s style. Crucially, DFJ’s largest gains occur on human-annotated benchmarks (DFJM-Human), fully independent of the bootstrapping pipeline, providing direct evidence of a general, human-aligned evaluation function rather than closed-loop imitation.

10. Training hyperparameters

We ran training on two AMD-MI210 GPUs, with 64 GB of memory each. DeepfakeJudge-7B took 4 hours to train, while DeepfakeJudge-3B took 2.5 hours. Table 8 shows the exact hyperparameters used while training our DeepfakeJudge models.

Hyperparameter	Value
Training	
Learning rate	1e-6
Train type	Lora
Lora Rank	32
Lora Alpha	64
Batch size (per device)	16
Gradient accumulation steps	1
Num. train epochs	2
Warmup ratio	0.05
Weight decay	0.0
Precision	bfloat16
Max sequence length	128,000
tune mm vision	False
tune mm mlp	True
tune mm llm	True
Evaluation	
max new tokens	2048
temperature	0.0

Table 8. Training and evaluation hyperparameters.

11. Inter-annotator Statistics for DeepfakeJudge-Meta-Human dataset.

Table 9 shows the pairwise and pointwise evaluation scores while preparing the DeepfakeJudge-Human dataset.

Metric	Value
Pairwise Evaluation	
Samples used	100 per annotator
Annotator 1 accuracy (vs. GT)	0.91
Annotator 2 accuracy (vs. GT)	0.95
Raw agreement (Annotator 1 & 2)	0.90
Cohen's κ	0.80
Both correct	88
Both wrong	2
Disagree, one correct	10
Pointwise Evaluation	
Samples used	100 per annotator
Annotator 1 MSE / RMSE	0.11 / 0.33
Annotator 2 MSE / RMSE	0.27 / 0.52
Inter-annotator MSE / RMSE	0.39 / 0.62
Annotator 1 exact match rate	0.89
Annotator 2 exact match rate	0.73
Pearson r (Annotator 1 / 2 / Overlap)	0.97 / 0.93 / 0.90
Spearman ρ (Annotator 1 / 2 / Overlap)	0.97 / 0.93 / 0.90
Both exact correct	65
One correct	31
Both wrong (same / different)	2 / 2

Table 9. Summary of human annotation reliability across pairwise and pointwise evaluations using annotators 1 and 2. Values are computed against automatically-derived ground truth.

Prompt filtering

Category / Class	Associated Keywords
people-portrait	portrait, headshot, close-up, candid, person, people, model, fashion, editorial, street portrait, wrinkles, freckles
people-activity	athlete, runner, dancer, musician, chef, worker, doctor, engineer, teacher, construction, artisan, farmer, baker
nature-landscape	mountain, valley, forest, desert, ocean, beach, river, waterfall, glacier, canyon, meadow, cliff, coast, island
animals-wildlife	wildlife, bird, eagle, owl, lion, tiger, bear, wolf, fox, deer, elephant, giraffe, zebra, whale, dolphin
animals-pets	dog, cat, puppy, kitten, hamster, parrot
architecture-exterior	building, skyscraper, bridge, street, alley, facade, historic, monument, castle, church, mosque, temple, pagoda, synagogue, stadium
architecture-interior	interior, living room, kitchen, bedroom, bathroom, office, workspace, studio, hotel lobby, museum, library, hallway
food-product	dish, meal, cuisine, plate, bowl, garnish, restaurant, pastry, bread, cake, coffee, espresso, latte, burger, pizza, sushi, product photo, packaging, bottle, label
transportation	car, train, bus, tram, bicycle, motorcycle, airplane, ship, harbor, terminal, station, highway, traffic, garage
sports-action	soccer, football, basketball, tennis, boxing, martial arts, skateboard, surfing, skiing, snowboarding, cycling, swimming
macro-detail	macro, close-up, texture, surface, dew, raindrops, insect, flower stamen
aerial-drone	aerial, drone, overhead, bird's-eye view, top-down
night-lowlight	night, low light, neon, long exposure, light trails, starry sky, milky way
weather	rain, snow, fog, storm, lightning, hail, mist, clouds
underwater	underwater, coral, reef, scuba, snorkel, sea turtle
events	wedding, festival, concert, parade, market, ceremony
fashion-beauty	runway, couture, makeup, hairstyle, wardrobe, studio portrait
documentary-street	street photography, documentary, photojournalism, everyday life, candid
unreal (negative)	dragon, fairy, wizard, elf, demon, vampire, alien, robot, cyberpunk, anime, cartoon, 3d render, digital art, surreal, concept art, illustration, ai-generated, futuristic, mythical, glowing, hologram, dreamscape
nsfw (negative)	nsfw, nude, explicit, erotic, fetish, sexual, breasts, lingerie, bondage, underage, gore, blood, decapitated, mutilated, NSFL

Table 10. Category and negative-class keyword sets used for prompt filtering. We evaluate the prompts from Wang et al. [45] using a weighted scoring function that emphasizes linguistic realism and photographic relevance. The final score combines three main components: prompt length (60% weight, modeled with a sigmoid favoring 30 to 100 words), clause count (30% weight), and a +0.5 bonus if any term from the photographic keyword whitelist appears. Additional penalties were applied for overly long or repetitive text. The resulting scores ranged from 0.70 to 1.38 (median ≈ 1.19), with higher values common since most prompts triggered the photo bonus. From 2,000 exported samples (columns: `prompt`, `category`, `score`), the average prompt length was about 55 words ($\sigma \approx 9$; range 45 to 122), and 84.9% of prompts contained explicit photographic terms. Thirteen categories were represented: `people-portrait` dominated with 1,697 prompts (84.9%), followed by `nature-landscape` (77), `transportation` (72), and smaller groups such as `animals-pets` (58) and `events` (15). Mean category scores varied, with `people-portrait` averaging 1.218 and others clustering around 0.71 to 0.77, reflecting the bias toward photographic realism.

Human Annotation

Flag Name	Check	Pass / Fail Description
Shadows	Do all shadows point roughly the same way, show a dark contact at touchpoints, and get softer as they stretch away?	PASS: All outdoor shadows lean the same way; tight dark line under shoes; edges soften outward. FAIL: One shadow points the opposite way; feet “float” (no dark contact); shadow edge equally sharp everywhere.
Lighting Match	Are objects lit from the same side with similar brightness/contrast for the scene?	PASS: Faces, walls, and props all brighter on the left side. FAIL: One object bright from the left while nearby objects are bright from the right.
Color Cast Consistency	Do whites/neutrals share the same warm/cool tint?	PASS: All whites slightly warm indoors. FAIL: A shirt is cold blue while everything else looks yellowish.
Relative Size / Scale	Do familiar things get smaller with distance and look believably sized?	PASS: Far people/cars noticeably smaller than near ones. FAIL: A far person is as big as someone close.
Perspective / Straight-Line Convergence	Do parallel lines (road edges, table sides) aim toward the same point(s)?	PASS: Building edges converge consistently. FAIL: An added sign’s edges converge to a different point.
Front-Back Overlap	Does the front object cleanly cover the back object without halos or see-through?	PASS: Vase cleanly blocks the book behind it. FAIL: Background leaks through hair/edges; color fringe halo.
Contact & Support	Do objects look supported and press into soft things when they should?	PASS: Cushion dips under a person; grass bends under a foot. FAIL: Heavy box sits on a blanket with zero dent.
Focus / Blur with Depth	Does blur increase smoothly with distance from the focus plane?	PASS: Foreground soft → subject sharp → background gently soft. FAIL: A pasted object stays razor-sharp while neighbors at the same depth are blurry.
Reflections & Transparency	Do reflections show the right thing at the right angle, and do transparent surfaces show both some reflection and some see-through?	PASS: Person appears (darker) in a shop window at the expected position; window shows faint outdoor reflection plus the room behind. FAIL: No reflection where expected, wrong pose/angle in mirror, or window behaves as a perfect mirror or perfectly clear at a glancing angle.
Material Shine	Do shiny things have crisp highlights and matte things have soft, broad highlights?	PASS: Metal = bright, sharp highlight; fabric = soft highlight. FAIL: Everything has the same plasticky shine.
Texture / Pattern Follow-Through	Do stripes, weaves, or wood grain bend and scale with folds and curves?	PASS: Shirt stripes curve smoothly around the sleeve. FAIL: Pattern stays flat or smeared across bends.
Text & Small Details	Are letters/numbers legible and small parts coherent?	PASS: Street sign letters readable, normal spacing. FAIL: Gibberish letters, melted numbers, warped watch face markings.
Object Completeness & Counts	Are obvious parts present and counted right (fingers, spokes, chair legs, petals, etc.)?	PASS: Hand shows five distinct fingers; bicycle has two wheels and believable spokes. FAIL: Wrong counts (e.g., three fingers), missing chair leg, duplicated petal, merged limbs.
Edges & Boundaries (cut-out / halo check)	Do object edges look like they naturally belong in the scene? Edges should be as sharp/soft as nearby stuff at the same distance, with no glowing outline, no weird color rim, and no “scissor-cut” look.	PASS: Jacket sleeve edge slightly soft like nearby edges; wispy hair/fur irregular (not clumped); no bright outline; edge grain matches background. FAIL: Bright/dark halo; too-straight cut line; color fringe; jaggy edge;

Table 11. Taxonomy of the flags used for human annotation

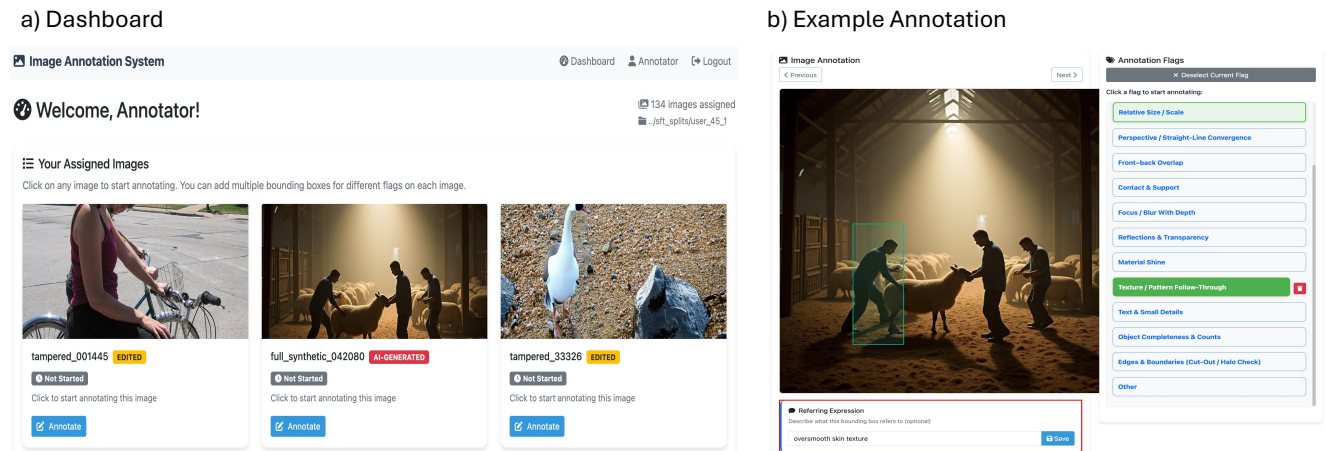


Figure 7. Overview of the human annotation framework. a) Shows the user dashboard, b) shows an example of the annotation process. The user selects a flag, draws the corresponding bounding box over the image, and then writes a referring expression.

Qualitative Examples for degraded reasoning responses.



(Image label: Real)

Rating	Reasoning
5 (Gold)	The bright sunlight casts a sharp shadow on the man's face, with clear shading under his nose and around his eyes. The scalp's highlights and soft skin texture show natural variation and pores. Background grass and water blur smoothly with depth, contrasting the sharp facial features in the foreground. The size and perspective of the man's head relative to the background landscape match a close-up view with distant elements.
4	The sunlight produces distinct shadows on the man's face, especially under his nose and near his eyes, which aligns with typical outdoor lighting. The scalp displays natural highlights, while the skin texture shows pores and slight variations. The background, including grass and water, has a smooth blur that matches the depth of field, complementing the sharpness of the face in the foreground. The head's size and perspective are consistent with the distant landscape, supporting the notion that this is an authentic close-up shot.
3	The image exhibits natural lighting, with shadows on the face and highlights on the scalp, although certain skin details seem a bit soft or slightly overexposed. The background blurs gently, consistent with a shallow depth of field, and the head size appears reasonable compared to the landscape. There are no apparent artifacts or edges that indicate manipulation, but the image lacks finer details in some regions.
2	While the face shows shadows and highlights, some skin areas appear unusually smooth or flat, hinting at potential retouching. The background blur is noticeable but inconsistent, with certain grass areas sharper than one would expect. Furthermore, the scale of the head seems slightly disproportionate to the distant scenery, and there are subtle blending artifacts around the ears. These details suggest that the image may not be entirely authentic.
1	The lighting on the face is not uniform, with shadows that do not correspond to the sunlight direction in the background. The scalp has an odd shine that resembles plastic, and the skin texture is devoid of natural pores. The background blur is patchy and inconsistent, and the head seems to be placed onto the landscape with unclear edges. These signs strongly indicate that the image is either fake or extensively edited.

Table 12. Example of a real image and corresponding degraded ratings produced by our bootstrapping process.



(Image Label: Edited)

Rating	Reasoning
5 (Gold)	The man in the center maintains a direct and intense gaze rather than lowering his eyes, which is unexpected given the context that suggests remorse or conflict. His expression lacks subtle emotional cues, such as softness or vulnerability, that might indicate regret. The overall posture and facial alignment seem slightly stiff, detracting from a natural, relatable interaction. The other individuals' expressions and body language appear more neutral, creating a contrast that highlights the central figure's unusual direct stare. This inconsistency in expression and gaze contributes to a sense of artificiality, suggesting the image may have been generated or manipulated.
4	The central figure's gaze is remarkably direct and confrontational, which is inconsistent with what one might anticipate from someone who is remorseful or anxious. His facial expression seems rigid and devoid of the emotional nuance found in authentic remorse. Meanwhile, the surrounding individuals exhibit more neutral and subdued expressions, which accentuates the central figure's unusual stare. This incongruity in emotional expressions and gaze suggests that the image may have been manipulated.
3	The man in the center looks directly at the camera with a strong gaze, which feels a bit unusual for the scene context. His face seems a little stiff compared to the others, who appear more relaxed or neutral. This difference makes the image feel less natural, though it's not fully clear if it's edited or not.
2	The eyes of the primary figure appear oddly intense, and the expression seems unnatural, while the others look fine. The direct gaze could suggest editing, but there are no clear signs of manipulation in other areas. It's possible this is an authentic image with just a peculiar expression or lighting that makes it look strange.
1	The man is looking straight at the camera and smiling, which is perfectly normal. There is no evidence of manipulation or synthetic content, and the expressions of the other people are in harmony with the environment. All elements seem authentic, with uniform lighting and natural stances.

Table 13. Example of an edited image and corresponding degraded ratings produced by our bootstrapping process.



(Image Label: Fake)

Rating	Reasoning
5 (Gold)	The skin on both girls looks overly smooth and lacks natural pores or texture, appearing unnaturally bright and plastic-like across their faces and hands. One of the fingers, particularly the left girl's hand in the lower center, is formed irregularly with an unnatural shape, disrupting the normal anatomy of the hand. The fingers seem merged or misshapen, breaking the expected geometry. Lighting on the skin is consistently even, which contrasts with the subtle shadows usually found in photographs. Additionally, the boundaries of the fingers and hands look somewhat blurred and unnatural compared to the crisp edges of the background elements. These flaws indicate the image is AI-generated.
4	The skin texture on the girls looks unnaturally smooth and bright, devoid of typical facial pores and natural skin variations. The left girl's hand, especially near the fingers, shows irregular shapes and merging that do not align with normal anatomy. The lighting is very uniform, lacking subtle natural shadows, and the edges around the fingers appear slightly blurred in contrast to the sharper background. These minor inconsistencies suggest that the image is probably AI-generated or heavily manipulated.
3	This image features both girls with skin that is unnaturally smooth, which looks somewhat unrealistic. Additionally, one finger on the left girl appears to be malformed or merged, not quite resembling a typical hand. The lighting is consistent and a bit flat, while the edges of the fingers are not very defined. These characteristics suggest that the image may have been digitally produced or manipulated, although some details remain unclear.
2	While the girls look largely authentic, certain parts of their skin seem overly smooth and bright, which might be due to lighting or editing. The left girl's fingers do not appear properly formed, possibly due to an awkward angle or blur. The overall lighting remains consistent, but the skin gives off a somewhat plastic-like impression. I am unsure if this image is genuine or artificial, but it appears to be edited.
1	The girls are happily holding snowmen and smiling. The lighting seems natural, and the bookshelves in the background are clear. The fingers and skin appear normal, lacking any obvious flaws or blurriness. This image resembles a standard photograph without evidence of manipulation or AI generation.

Table 14. Example of a fake image and corresponding degraded ratings produced by our bootstrapping process.



(Image Label: Edited)

Rating	Reasoning
5 (Gold)	The eyes of the man on the right appear unnaturally warped and lack realistic detail, contributing to a disturbing facial expression. The intersection between the spectacle and the face of the man on the left shows unnatural blending, with parts of the glasses and skin merging in an unrealistic way. This overlap issue creates an impression of transparency rather than proper layering. The lighting and edges around the glasses and facial features also look inconsistent, causing a blurred boundary effect. These distorted eyes and implausible overlaps are clear signs that the image was generated by AI.
4	The individual on the right has eyes that are distinctly distorted with unnatural shapes, and his mouth is overly open, which looks unrealistic. Moreover, the spectacles on the left man's face do not blend effectively; the edges around the glasses seem to leak into the skin, causing an unusual transparency effect. These overlapping artifacts and the irregularity of the eyes suggest that the image is either AI-generated or has been heavily manipulated.
3	The person on the right has eyes that look strange and unrealistic, and the glasses on the left seem to blend improperly with the face. The lighting around the glasses and the edges of the face is uneven, which makes the image seem somewhat unnatural. These elements suggest possible manipulation or synthetic modification, even if the exact problems are not distinctly outlined.
2	The eyes on the right appear unusual, possibly due to the expression or lighting conditions. The glasses on the left individual seem somewhat unclear where they connect to the face, which might indicate a quality issue with the photograph. While there are signs of potential editing, the overall scene looks relatively normal.
1	The image appears generally normal; the eyes and glasses look appropriate without noticeable distortions or blending issues. The lighting is consistent across the faces and objects in the scene. It resembles a casual photograph with no indications of manipulation or artificiality.

Table 15. Example of an edited image and corresponding degraded ratings produced by our bootstrapping process.



(Image Label: Fake)

Rating	Reasoning
5 (Gold)	The train in the mid-right appears unnaturally bright and overly smooth, lacking realistic surface textures that would normally show wear and detail. The entire upper-right train section is distorted, with shapes fused together and unclear details, making it hard to distinguish individual parts. The merging of elements creates an unnatural and confusing structure that does not align with typical train forms. Edges between the train and the background are uneven and lack depth, further weakening the image's realism. These texture and structural anomalies clearly indicate that the scene is artificially generated.
4	The train located on the right side is strikingly bright and shows a lack of detailed texture, giving it an overly smooth and synthetic appearance. In the upper-right section, the shapes seem to merge, complicating the identification of distinct train parts. The boundaries between the train and its background are somewhat uneven, further diminishing its natural look. These texture irregularities and the blending of forms strongly suggest that the image is AI-generated.
3	The train exhibits an unusually bright and smooth texture, with certain sections blending into one another in the upper right area. The distinct features of the train's structure are somewhat obscured, and the edges around it lack sharpness, indicating possible alterations. Overall, the image raises suspicion and does not appear entirely authentic, despite some visible details.
2	The train looks a little off because some parts seem merged and unclear, especially on the upper right. The colors also look a bit unnatural. However, the rest of the scene looks okay. It might be edited or just a low-quality photo. There are no strong clear signs, so I'm not completely sure.
1	The train looks normal and the tracks appear fine, with no visible issues. The colors and textures seem consistent with a real photo. The background and surrounding objects also look natural. Nothing stands out as fake or edited here, so this image is definitely real.

Table 16. Example of fake image and corresponding degraded ratings produced by our bootstrapping process..



(Image label: Real)

Rating	Reasoning
5 (Gold)	Soft, natural daylight uniformly illuminates the face and the knit hat, creating subtle shadows beneath the eyebrows and nose that correspond with the light source. The knitting texture on the hat is consistent in scale and shows realistic stitches and shadows. The face's sharpness contrasts smoothly with the slight blur of the background elements, providing depth. The hair near the hat edges blends naturally without halos, and the skin tones show natural variation in hue and lighting.
4	The lighting is gentle and natural, providing even daylight across both the face and the knit hat. Shadows beneath the eyebrows and nose align appropriately with the light source's direction. The knitting texture on the hat displays clear, realistic stitch details and a consistent scale. The face is sharply focused, contrasting with the slightly blurred background, which adds depth. The hair edges adjacent to the hat blend naturally without visible halos. Skin tones show credible variations in color and lighting, supporting the image's authenticity.
3	The image shows a daylight scenario that generally looks natural, with mild shadows present on the face and hat. The knit texture of the hat is apparent, though some areas lack detail. The face is sharply focused while the background is somewhat blurred, enhancing the depth. Hair along the edges of the hat blends reasonably well, although a few spots could be smoother. Skin tones appear normal, but subtle inconsistencies might stem from lighting or minor edits. Overall, the image seems real, though not completely flawless.
2	The illumination on the face and hat is uneven, with certain shadows looking inconsistent or too sharp for natural daylight. The knit texture on the hat is visible but appears somewhat flat in certain areas, suggesting possible retouching. The background blur is irregular, and some hair edges near the hat seem slightly unnatural or haloed. Skin tones vary, but some regions look excessively smoothed or edited. These factors cast doubt on the image's complete authenticity, though it may not be heavily altered.
1	The lighting is mismatched, with shadows on the face going in different directions and inconsistent with the outdoor background. The knit hat texture looks artificial and overly uniform, lacking natural stitch variation. Hair edges around the hat show clear halos and blending errors. The background blur is uneven, and skin tones appear overly airbrushed or colored unnaturally. These problems strongly indicate the image has been digitally manipulated or is fake.

Table 17. Example of a real image and corresponding degraded ratings produced by our bootstrapping process..

Qualitative Examples of Paraphrased Gold-Standard Ratings.



(Image Label: Edited)

Variant Reasoning

- Gold The man's face and head area exhibit unnatural texture, appearing synthetic and inconsistent with typical skin and hair details. Additionally, the motorcycle seat shows a dip as if supporting weight, yet the man is seated further forward, not making contact with that indentation. These discrepancies in texture and physical contact indicate manipulation in these areas. The rest of the image, including the motorcycle, background, and rider's clothing, appears visually consistent and realistic.
- Para-1 The man's face texture looks somewhat unnatural, with an inconsistent pattern that suggests synthetic alteration. Also, the motorcycle seat shows a clear dip as if bearing weight, but the man is seated slightly forward, not pressing down on that area. These visual clues suggest the face region was edited. The rest of the scene, including the motorcycle and background, seems consistent and natural.
- Para-2 The texture of the man's face appears somewhat artificial, exhibiting an inconsistent pattern that indicates possible synthetic modification. Additionally, the motorcycle seat shows a noticeable dip as if someone is applying pressure, yet the man is positioned slightly forward, not affecting that area. These visual indicators imply that the face region has been altered. The remainder of the scene, including the motorcycle and background, appears coherent and natural.
- Para-3 The texture on the man's face seems oddly synthetic, with a pattern that lacks consistency, hinting at digital alteration. Furthermore, the motorcycle seat exhibits a distinct depression, suggesting weight is applied, while the man is seated slightly forward, not influencing that section. These visual signs suggest that the facial area has been modified. The rest of the scene, such as the motorcycle and background, looks consistent and realistic.
- Para-4 The man's facial texture appears somewhat unnatural, with a pattern that suggests it may have been synthetically altered. Moreover, the motorcycle seat displays a clear indentation, implying that weight is being exerted, yet the man is leaning slightly forward, not applying pressure to that area. These visual cues indicate that the face region has undergone editing. The rest of the scene, including the motorcycle and background, seems consistent and authentic.
-

Table 18. Comparison of a fake image and the corresponding paraphrased and Gold-standard reasoning produced by our bootstrapping process.

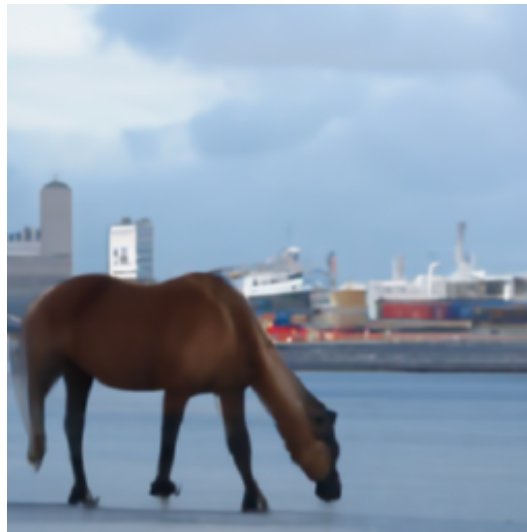


(Image Label: Fake)

Variant Reasoning

- Gold** Two animals blend confusingly into the surrounding foliage, lacking clear form or recognition, with edges smudged against leaves and branches. The front leaves exhibit inconsistent lighting compared to the overall scene, appearing artificially bright with an uneven glow that doesn't match the ambient light direction. Shadows under the animals are absent, resulting in unnatural floating effect and weak grounding. The perspective appears compressed in the lower-right, making the background look too close and flattened, disrupting spatial depth cues. These issues with object blending, lighting disparity, missing shadows, and distorted depth strongly indicate the image is AI-generated.
- Para-1** The image shows two animals poorly defined and blending into the dense foliage, with unclear boundaries that reduce recognizability. The lighting on the leaves in the foreground is uneven and doesn't align well with the rest of the scene, suggesting inconsistent illumination. Notably, shadows beneath the animals are missing, causing them to appear as if floating above the ground. The perspective in the lower-right section looks unnaturally compressed, making the background seem too near and flattening the spatial depth. These subtle inconsistencies in object clarity, lighting, shadows, and perspective point toward an artificially generated image.
- Para-2** The image depicts two animals that are not clearly defined and are blending into the surrounding foliage, resulting in indistinct edges that hinder identification. The lighting on the foreground leaves is inconsistent and does not match the overall scene, indicating uneven illumination. Additionally, the shadows beneath the animals are absent, making them appear as if they are floating. The perspective in the lower-right area appears unnaturally compressed, causing the background to seem too close and flattening the depth of the scene. These subtle discrepancies in object definition, lighting, shadows, and perspective suggest that the image is artificially generated.
- Para-3** This image features two animals that are poorly defined and merge with the dense foliage, leading to unclear boundaries that make recognition difficult. The lighting on the leaves in the foreground is uneven and does not correspond well with the rest of the scene, suggesting inconsistency in illumination. Notably, the absence of shadows under the animals creates an illusion of them floating above the ground. The perspective in the lower-right section appears unnaturally squished, making the background look too close and diminishing spatial depth. These minor inconsistencies in object clarity, lighting, shadows, and perspective indicate that the image is likely artificially generated.
- Para-4** The image shows two animals that are not distinctly visible and blend into the thick foliage, resulting in unclear outlines that impede recognition. The lighting on the leaves in the foreground is inconsistent and does not align with the overall scene, hinting at uneven illumination. Importantly, the shadows beneath the animals are missing, giving them a floating appearance. The perspective in the lower-right corner seems unnaturally compressed, causing the background to appear too close and flattening the spatial depth. These subtle inconsistencies in object definition, lighting, shadows, and perspective suggest that this image is artificially generated.
-

Table 19. Comparison of a fake image and the corresponding paraphrased and Gold-standard reasoning produced by our bootstrapping process.



(Image Label: Fake)

Variant Reasoning

Gold	The horse appears to be missing its left hind leg, resulting in a strange, incomplete posture. Its ears and face display inconsistent shapes and edges that do not align naturally, lending a distorted appearance. Despite the reflective surface of the water beneath, the horse's reflection is entirely absent, which breaks expected visual continuity. Additionally, the horse seems to be walking directly on the water's surface rather than on solid ground, creating an unrealistic effect of suspension. These geometric inconsistencies, lack of proper reflections, and object incompleteness strongly suggest that the image was AI-generated.
Para-1	The horse's left hind leg is notably absent, causing an unnatural pose. The ears and face show some irregularities in shape that suggest digital alteration. The reflection on the water is missing where it should appear, breaking the expected visual consistency. Also, the horse appears to stand on water rather than solid ground, which is unrealistic. These signs of incomplete objects and missing reflections strongly indicate AI generation.
Para-2	The left hind leg of the horse is conspicuously missing, resulting in an awkward stance. Additionally, the shape of the ears and face presents inconsistencies that imply digital manipulation. The expected reflection on the water is absent, disrupting the visual coherence. Furthermore, the horse appears to be walking on water, which is implausible. These indicators of incomplete elements and absent reflections strongly suggest AI generation.
Para-3	The horse's left hind leg is clearly missing, leading to an unnatural posture. There are also discrepancies in the shape of the ears and face that hint at digital alterations. The reflection on the water is lacking, which undermines the visual consistency. Moreover, the horse seems to be standing on water instead of solid ground, which is unrealistic. These signs of incomplete objects and missing reflections strongly indicate AI generation.
Para-4	The absence of the horse's left hind leg is striking, resulting in an unnatural pose. The ears and face exhibit irregular shapes that suggest they have been digitally altered. The reflection on the water is missing where it should be, disrupting the expected visual harmony. Additionally, the horse appears to be walking on water, which is not realistic. These indicators of incomplete objects and absent reflections strongly imply AI generation.

Table 20. Comparison of a fake image and the corresponding paraphrased and Gold-standard reasoning produced by our bootstrapping process.

Prompts

Prompt Template for Degraded Reasoning Generation

You are provided with an image, the ground truth classification (real/fake/edited), a description about the reasoning rated 5 (perfectly accurate, highest quality), a JSON list showing the human annotation on the image (raised flags, bounding boxes normalized to 1-1000, and referring expressions), and a flag taxonomy of human-familiar forensic cues used by annotators. Your task is to generate four additional responses that simulate progressively lower-quality reasoning outputs for the same image. Each generated response should correspond to a quality rating from 4 to 1, where Rating 5 is the provided gold standard and Ratings 4 through 1 represent decreasing quality. As the rating decreases, the responses should reflect increasing levels of degradation (for example, wrong classification, hallucinations, omissions, irrelevant details, logical inconsistencies, misused or missed flags, grammatical issues, or wording issues). Only do wrong classification for very low ratings. All generated responses must be similar in length to the gold standard and maintain the expected task format (for example, short answer plus explanation). Keep in mind that for real images there is no human annotation; the gold standard is based on the absence of certain random flags. When misclassifying, misclassify between the three classes randomly (real <-> fake <-> edited). For edited images, the label also contains the bounding box for the fake region, so use it when creating lower-rating reasoning outputs when necessary. Do not simply truncate the gold response. Use the image and human annotation to ground correctness; vary style and wording naturally. Make sure that the answers are degraded based on human perception. The higher rated answers are more useful to humans for understanding the outputs.

Task:

Generate four degraded responses corresponding to quality ratings 4 through 1, based on the gold standard response (Rating 5). Follow the rating guidelines strictly.

Rating Guidelines:

- Rating 4 - Mostly accurate with minor issues. Preserves key meaning; small lapses in detail or precision; flags mostly correct; clear wording, random wrong flag added/removed.
- Rating 3 - Partially correct. Conveys the general idea but with noticeable errors, omissions, or misinterpretations; some generic phrasing; flags partly correct or incomplete; clarity or helpfulness mixed; unclear wording; sometimes underconfident tone but correct classification.
- Rating 2 - Poor alignment (may misclassify). Serious flaws; weak grounding; mixes valid and invalid details; flags often wrong or irrelevant; vague or hedged language; sometimes underconfident tone but wrong classification.
- Rating 1 - Unrelated or incorrect (often misclassifies). Fails to reflect the image or instruction; hallucinated content; misleading or nonsensical flags; confusing wording; sometimes high-confidence tone but wrong classification.

Output Format:

Return a JSON object with exactly the keys "rating_4", "rating_3", "rating_2", and "rating_1". Each value must be only the generated response text for that quality rating (no extra fields).

Flag Taxonomy:

{Flag descriptions}

EXAMPLE:

Image:

See the first attached image.

Label:

real

Human Annotation:

{}

Input:

Image:

{Image}

Label (contains the edited region for edited class):

{Label}

Human Annotation (only for fake and edited):

{Human annotation}

Gold Standard Response (Rating 5):

{gold_standard_response}

Output (according to the JSON schema):

```
{
  "rating_4": "[Generated response]",
  "rating_3": "[Generated response]",
  "rating_2": "[Generated response]",
  "rating_1": "[Generated response]"
}
```

Prompt Template for Evaluator model.

You are provided with an image, the ground truth classification (real / fake / edited), a gold-standard reasoning rated 5 (perfectly accurate, highest quality), a JSON list of human annotations for the image (raised flags, any bounding boxes normalized to 1-1000, and referring expressions), a flag taxonomy of human-familiar forensic cues, and a set of candidate responses (each typically containing a short answer plus an explanation).

Your task is to evaluate each candidate response and assign an evaluation rating from 1 to 4. While the gold-standard response (Rating 5) is a reference for what ideal reasoning looks like, your evaluation should primarily judge each candidate on how well it fulfills the instruction for this image, grounded in the image content, label, and human annotations. Focus your reasoning on what is incorrect, missing, misleading, or poorly justified in the candidate response itself.

Special notes and constraints:

- For real images, the human annotation is typically empty; the gold reasoning reflects the absence of artifact flags. Penalize candidates that invent edited regions, hallucinate artifacts, or misuse flags without evidence.
- For fake (fully generated/synthetic) images, human annotations may include raised flags but no edited-region bounding box. Evaluate whether the candidate's flags match plausible generative artifacts and whether claims are grounded in visible cues.
- For edited images, the label includes the edited region bounding box(es). Evaluate whether candidates correctly identify the manipulation and localize it (if they attempt localization). If a candidate provides a box or a referring expression, check it against the provided edited region(s) for plausibility (qualitatively; coarse overlap is acceptable). Missing or obviously wrong localization is a major error.
- Do not reward or penalize writing style alone; judge clarity insofar as it affects usefulness and correctness. Avoid over-reliance on the gold wording, evaluate the content and grounding.
- Treat the flag taxonomy below as the authoritative list of cues. Using non-existent flags, or misnaming/misapplying flags, should be penalized proportionally.

Evaluation Steps

1. Read the instruction to understand the expected output format (e.g., "short answer + explanation") and task focus (classification, localization, reasoning).
2. Use the image + label + human annotations (and the edited-region box if applicable) as the factual basis.
3. Refer to the gold-standard response (Rating 5) only as a reference for correctness and completeness, not as wording to match.
4. For each candidate response, identify:
 - Classification correctness (does the short answer match the label?).
 - Grounding quality (are cited cues visible and consistent with the taxonomy?).
 - Proper use (or misuse) of flags, localization, sizes/coordinates, and referring expressions.
 - Hallucinations, contradictions, irrelevancies, omissions, or logical gaps.
5. Assign a rating from 1-4 to each candidate.
6. Write a brief rationale for each rating, pointing to specific strengths/weaknesses (e.g., correct flags but missed localization; misclassification; hallucinated halos; misapplied perspective).

Rating Guidelines (Evaluation)

Rating Guidelines:

- Rating 4 - Mostly accurate with minor issues. Preserves key meaning; small lapses in detail or precision; flags mostly correct; clear wording, random wrong flag added/removed.
- Rating 3 - Partially correct. Conveys the general idea but with noticeable errors, omissions, or misinterpretations; some generic phrasing; flags partly correct or incomplete; clarity or helpfulness mixed; unclear wording; sometimes underconfident tone but correct classification.
- Rating 2 - Poor alignment (may misclassify). Serious flaws; weak grounding; mixes valid and invalid details; flags often wrong or irrelevant; vague or hedged language; sometimes underconfident tone but wrong classification.
- Rating 1 - Unrelated or incorrect (often misclassifies). Fails to reflect the image or instruction; hallucinated content; misleading or nonsensical flags; confusing wording; sometimes high-confidence tone but wrong classification.

Output Format

Return only a JSON object whose keys exactly mirror the candidate identifiers provided in Generated Responses (e.g., "candidate-1", "candidate-2",...). Each value must be a JSON object with exactly these keys:

- "rating": an integer in {1, 2, 3, 4}
- "rationale": a concise string explaining the rating (no more than a few sentences)

Do not include any additional keys or text outside the JSON object.

Flag Taxonomy (Reference for Evaluation)
{Flags}

EXAMPLE

Image: See the first attached image.

Label: real

Human Annotation: {}

Output Format:

```
{"candidate_1": {"rating": <1-4>, "rationale": "<concise explanation>"}, "...": {...}}
```

Input:

Image:

See attached image.

Label:

{label_text}

Human Annotation:

{human_annotation}

Gold Standard Response (Rating 5):

{gold_standard_response}

Generated Responses:

{generated_responses}

Output (according to the JSON schema):

{output_format}

Prompt Template for Regeneration.

You are provided with an image, an instruction, the ground-truth classification (real / fake / edited), a gold standard response rated 5 (perfectly accurate, highest quality), a JSON list of human annotations for the image (raised flags, any edited-region bounding boxes normalized to 1-1000, and referring expressions), a flag taxonomy of human-familiar forensic cues, and a set of generated responses, each intended to match a specific quality rating from 4 to 1. However, some generated responses were evaluated and found to deviate from their intended quality levels.

Task:

Your task is to regenerate revised responses only for those entries where the absolute difference between the intended rating and the evaluation rating is greater than zero (i.e., $|\text{intended rating} - \text{eval rating}| > 0$). For each such entry, produce a revised response that strictly conforms to the intended quality rating, based on the definitions in the rating guidelines below. Use the image and human annotations (and edited-region box if applicable) as the factual grounding for determining what content is valid. Adjust the response so that its evaluation rating would now exactly match the intended rating.

- If the eval rating is higher than the intended rating, degrade the response by introducing errors such as mild hallucinations, factual distortions, omissions, vagueness, or grammar issues (consistent with the target rating).
- If the eval rating is lower than the intended rating, improve the response by clarifying cues, reducing errors, restoring key context from the annotations/taxonomy, and aligning tightly with the label and visible evidence.

Rating Guidelines:

- Rating 4 - Mostly accurate with minor issues. Preserves key meaning; small lapses in detail or precision; flags mostly correct; clear wording.
- Rating 3 - Partially correct. Conveys the general idea but with noticeable errors, omissions, or misinterpretations; some generic phrasing; flags partly correct or incomplete; clarity or helpfulness mixed; underconfident tone but correct classification.
- Rating 2 - Poor alignment (may misclassify). Serious flaws; weak grounding; mixes valid and invalid details; flags often wrong or irrelevant; vague or hedged language; underconfident tone but wrong classification.
- Rating 1 - Unrelated or incorrect (often misclassifies). Fails to reflect the image or instruction; hallucinated content; misleading or nonsensical flags; confusing wording, high-confidence tone but wrong classification.

Flag Taxonomy :

{Flags description}

Example:

Image: See the first attached image.

Label: real

Human Annotation: {}

{}

Output Format:

Your final output must be a valid JSON object. For each entry where $|\text{intended rating} - \text{eval rating}| > 0$, include a key "rating {i}" and update the value with a newly revised response that adheres precisely to the intended quality rating. If the eval rating is equal to the intended rating, do not modify or include that entry. Only output the revised responses.

Input:

Image:

{image}

Label (real / fake / edited):

{label}

Human Annotation (JSON with raised flags, boxes 1-1000, referring expressions):

{human_annotation}

Gold Standard Response (Rating 5):

{gold_standard_response}

Feedback Data (JSON):

{feedback_data}

Output (according to the JSON schema):

{output_format}

Prompt Template for Rephrasing.

You are given context about an image-authenticity reasoning task.

Instruction:

{instruction}

Label: {label}

Human Annotation: {human_annotation}

Original response (to paraphrase):

{response_text}

Task:

Generate four paraphrases that preserve every factual claim, label decision, flag usage, bounding box reference, and XML-style tag structure (e.g., `<reasoning>...</reasoning>`, `<answer>...</answer>`). Each paraphrase should be similar in length to the original but use distinct wording. Do NOT introduce new information, remove evidence, or change the classification. Retain the same overall format and tags.

Output only a JSON object: `{{"paraphrase_1": "...", "paraphrase_2": "...", "paraphrase_3": "...", "paraphrase_4": "..."}}}`

Prompt Template for Rephrasing.

You are given context about an image-authenticity reasoning task.

Instruction:
{instruction}

Label: {label}
Human Annotation: {human_annotation}

Original response (to paraphrase):
{response_text}

Task:

Generate four paraphrases that preserve every factual claim, label decision, flag usage, bounding box reference, and XML-style tag structure (e.g., <reasoning>...</reasoning>, <answer>...</answer>). Each paraphrase should be similar in length to the original but use distinct wording. Do NOT introduce new information, remove evidence, or change the classification. Retain the same overall format and tags.

Output only a JSON object: {"paraphrase_1": "...", "paraphrase_2": "...", "paraphrase_3": "...", "paraphrase_4": "..."}"

Prompt Template for pointwise model evaluation.

You are provided with a image, a corresponding ground truth class (real/fake/edited), and a response generated by a model. The response consists of the model's final prediction, along with a rationale describe how the model reached that decision.

Your task is to evaluate the quality of the response, considering how well it explains the classification behind its prediction, based on the content of the image and the ground truth. This is a holistic judgment and should be based on the overall correctness, relevance, completeness, and grounding of the response.

Task:

For each response:
- Assess how well it describes the reasoning behind the prediction, and how well it is grounded in the image.
- Consider whether the response is accurate, relevant, complete, and grounded in the image. For edited images, also use the bounding box information for rating the responses.
- Provide a brief rationale explaining the overall quality and alignment of the response inside <reasoning></reasoning>.
- Output a score from 1 (worst) to 5 (best) indicating the overall quality inside <score></score>.

Rating Guidelines:

- Rating 5 - Fully accurate, complete, and well-grounded. The response precisely explain the reason behind the correct classification with no notable errors or omissions.
- Rating 4 - Mostly accurate with minor issues. Preserves key meaning; small lapses in detail or precision; flags mostly correct; clear wording, random wrong flag added/removed.
- Rating 3 - Partially correct. Conveys the general idea but with noticeable errors, omissions, or misinterpretations; some generic phrasing; flags partly correct or incomplete; clarity or helpfulness mixed; unclear wording; sometimes underconfident tone but correct classification.
- Rating 2 - Poor alignment. Serious flaws; weak grounding; mixes valid and invalid details; flags often wrong or irrelevant; vague or hedged language; sometimes underconfident tone but wrong classification.
- Rating 1 - Unrelated or incorrect. Fails to reflect the image or instruction; hallucinated content; misleading or nonsensical flags; confusing wording; sometimes high-confidence tone but wrong classification.

Input:

Image: Check the attached image

Ground Truth Label: {label}

Candidate response:

{candidate_response}

Output (Strict Format):

<reasoning>{{your reasoning and explanation for the rating}}</reasoning>
<score>{{integer score from 1 to 5}}</score>

Prompt Template for pairwise model evaluation.

You are provided with an image, a corresponding ground-truth class (real/fake/edited), and two responses generated by different models. Each response consists of the model's final prediction plus a rationale.

Your task is to compare the two responses and decide which one better aligns with the ground truth classification, based on the content of the image. Make a holistic judgment that considers correctness, relevance, completeness, and grounding.

Task:

For the given pair of responses:

- Judge which response is better reasoning for the classification
- Consider whether each response is accurate, relevant, complete, and grounded in the video.
- Output only A or B, wrapped strictly inside <answer></answer> tags.

Evaluation Guidelines:

- Accuracy: Prefer responses that are factually correct and consistent with the image, and have the correct classification
- Relevance: Prefer responses that directly mention the reasoning behind the classification
- Completeness: Prefer responses that capture all key aspects needed for a full grounded answer, containing the exact issues/non-issues with the image.
- Grounding: Prefer responses clearly supported by the image, avoiding hallucinations.
- If one response contains hallucinations, irrelevant content, or omissions, prefer the other.
- If both responses are strong, choose the one that is more precise and detailed.
- If both responses are weak, choose the one that is less flawed.

Ground Truth Label: {label}

Input:

Response A:
{response_a}

Response B:
{response_b}

Output (Strict Format):

<answer>{{A_or_B}}</answer>

Prompt Template for generating gold standard rating from human-annotated fake data.

You will receive: (1) an image and (2) a JSON-like string named "flags" listing issues. Each entry includes "bboxes": [{"coordinates": [x1,y1,x2,y2], "ref_exp": "..."}] with coordinates normalized 1-1000; entries may include "timestamp" (ignore it).

USE OF FLAGS
Treat the entries as hints about WHAT and WHY specific regions look synthetic. Never mention JSON, flags, coordinates, boxes, or timestamps in your writing. Write as if you are simply looking at the image and grounding your observations in visible evidence consistent with the hints. Convert some regions into coarse spatial phrases (upper-left, upper-center, upper-right, mid-left, center, mid-right, lower-left, lower-center, lower-right) only when the SAME anomaly occurs multiple times and needs differentiation; otherwise, omit location phrases and refer to the objects directly. Ignore the referring expression if it does not entail a flaw in the image, but rather contain instructions about things that might have been done to the image

OUTPUT FORMAT (exactly two lines, nothing else):
<think>
Write 4-6 sentences (about 70-120 words) describing concrete observations. Integrate and paraphrase the hinted issues naturally (e.g., "skin looks unnaturally smooth," "hands are brighter than nearby skin," "faces are distorted"). Use at least FOUR distinct cues overall: one texture/material cue, one lighting/color cue, one geometry/structure cue, and one boundary/depth cue. Conclude that these problems indicate the image is AI-generated.
</think>
<answer>fake</answer>

RULES
Neutral, observational tone; do not identify real people.
Vary phrasing while staying faithful to the visible evidence and the hinted issues.

Prompt Template for generating gold standard rating from real data.

You are an image-authenticity checker. For this task the ground-truth label is Real. Keep this private. Your job is to produce a short, observation-based justification that naturally supports a Real verdict.

OUTPUT FORMAT (exactly two lines, nothing else):
<think>[4-7 sentences; 70-130 words of observations]</think>
<answer>Real

DELIVERABLE RULES
Base the <think> text only on what is visible in the image. Refer to concrete objects (e.g., "plate," "cherries," "helmet," "gravel," "glasses"). Use at least FOUR distinct visual cues, including at least:
- one photometric cue (shadows, lighting, or color cast),
- one geometric cue (scale or perspective/line convergence),
- one boundary/depth cue (contact/support, overlap, focus/blur, or edges/halo).
Write in present tense, neutral and concise; avoid purple prose, probabilities, or self-references. Do NOT mention any "checkboxlist," "flags," or that you were given the label. Do NOT identify real people; describing non-identifying attributes (e.g., clothing, pose) is fine. No bullet lists or newlines inside <think>; use sentences only. The second line must read exactly: <answer>Real</answer>

CUE LIBRARY (use, but do not name explicitly)
1) Shadows - Directions align; dark contact at touchpoints; edges soften with distance.
2) Lighting Match - Same light side/contrast across scene.
3) Color Cast Consistency - Whites/neutrals share similar warm/cool tint.
4) Relative Size / Scale - Familiar objects are plausibly sized and diminish with distance.
5) Perspective / Line Convergence - Parallel edges aim to consistent vanishing point(s).
6) Front-Back Overlap - Near objects cleanly occlude far ones; no halos/see-through.
7) Contact & Support - Objects rest firmly; soft materials compress; no "floating."
8) Focus / Blur with Depth - Sharpness changes smoothly with distance; neighbors at same depth share sharpness.
9) Reflections & Transparency - Reflections show correct content/angle; glass shows some reflection and some see-through.
10) Material Shine - Shiny = crisp highlights; matte = broad/soft highlights.
11) Texture / Pattern Follow-Through - Stripes/grain bend and scale with form/folds.
12) Text & Small Details - Letters legible; tiny parts coherent.
13) Object Completeness & Counts - Obvious parts present and counted correctly.
14) Edges & Boundaries - Edge sharpness matches surroundings; no glow, color fringing, or scissor-cut look.
15) Other - Any additional visual consistency that supports realism.

QUALITY NOTES
Vary which cues you emphasize across images to avoid repetitive language. Always tie observations to specific objects and spatial relations rather than abstract claims.
Remember: exactly two lines of output, with the second line fixed to <answer>Real</answer>

12. Out-of-Distribution Generalization of DeepfakeJudge Model

To demonstrate that our models, DeepfakeJudge-3B and DeepfakeJudge-7B, can effectively evaluate deepfake reasoning on data outside our training pipeline, we assess them using the ground truth human-annotated reasonings from the DD-VQA dataset [54], which contains reasonings for face-swapped images. We convert the test set to our input format, placing the reasoning within `<reasoning>...</reasoning>` and the prediction inside `<answer>...</answer>`. The model then generates a rating along with an explanatory rationale. On average, our model assigns a rating of **3.18** across all responses. Table 21 presents several qualitative examples of these ratings and the corresponding rationales. The rationales produced by our model are clear, informative, and well-grounded in the visual content. Moreover, these rationales provide valuable interpretability and can serve as reliable signals for large-scale automatic supervision or self-evaluation pipelines in future deepfake detection systems.

Table 21. Rationales and scores generated by DeepfakeJudge-3B on the ground truth reasonings of the test-set of DD-VQA. DeepfakeJudge-3B correctly identifies where the reasoning is ungrounded and superficial, and gives a rating accordingly.

Image	Ground Truth Reasoning (from DDVQA)	Evaluation Response	Score
	<p>The image looks fake. It is an image with manipulated face regions. The person's mouth looks a bit fake because overly small mouth underage individuals. The person's skin looks a bit fake because the skin is overly smooth and lacks of details, such as skin texture the person has mismatched bangs.</p>	<p>This response correctly identifies the manipulated face regions and mentions the smooth skin, which aligns with the ground truth label. However, it incorrectly suggests the person is underage, which is not supported by the image.</p>	3
	<p>The image looks fake. It looks like a computer-generated artist styled image for animation, carton or films. The person's eyes look very fake because the eyes look unnaturally asymmetric. The person has mismatched bangs. The person's nose looks very fake because the nose looks curved.</p>	<p>This response correctly identifies the image as fake, noting asymmetry and mismatched features. However, it lacks specificity about the style and suggests the person is animated, which is not entirely accurate.</p>	3
	<p>The image looks fake. It looks like an ai-generated image although with human appearance but with unrealistic texture or lighting. The person's eyes look a bit fake because the eyes look unnaturally asymmetric. The eyes looks too rigid and lifeless to convey expression. The person has mismatched bangs. Mismatched mustache. The lighting in the image is inconsistent, unrealistic shadows. The person's nose looks very fake because the nose looks curved. The nose is placed at the wrong place on the face.</p>	<p>Fully accurate, complete, and well-grounded. The response precisely explain the reason behind the correct classification with no notable errors or omissions.</p>	5
	<p>The image looks fake. It is an image with manipulated face regions. The person's skin looks very fake because the skin is overly smooth and lacks of details, such as skin texture there are stains/flaws on the person's face there are areas looking unnaturally brighter there area of looking unnaturally darker the person's nose looks a bit fake because the nose do not align properly with other facial features. The person's eyebrows look a bit fake because the eyebrows do not match the face's structure.</p>	<p>This response accurately identifies the manipulation of facial features, noting the lack of natural skin texture and the unnatural brightness and darkness. However, it could have been more specific about the alignment issues with the nose and eyebrows.</p>	4