

---

## Appendix for: Understanding Generative AI Capabilities in Everyday Image Editing Tasks

---

### A. Dataset Statistics

We collect 98,234 posts and 350,609 edited images from Reddit, with 58,624 posts sourced from PushShift between 2013 and 2022 and 39,610 gathered between October 2024 and early 2025. After processing, the final dataset includes 82,976 posts and 305,806 edited images, with 71,027 single-image and 11,949 multiple-image requests. Tab. A1 shows the breakdown of the dataset composition, and Figure A5 illustrates key statistics and trends within the data.

Our dataset contains a diverse set of subjects, with 49,134 unique subjects across all requests. The *People and Related* category is the most common, accounting for 53.5% of the requests (Tab. A2). Subject trends exhibit seasonal variations (Fig. A4), with family-related requests peaking around the holiday season and New Year. Additionally, Fig. A2 illustrates the most common subjects.

The most common action is *delete*, accounting for 32.9% of all edits (Fig. A3), a trend that has remained dominant over the years (Fig. A5a). These involve removing *people*, such as photobombers, for aesthetic or personal reasons. Other deletion requests involve removing *objects*, like poles, bags, or signs, in order to reduce visual clutter as well as eliminating *facial imperfections*, such as acne or wrinkles for social media or professional use. The most frequent types of actions applied to subjects are illustrated in Fig. A5b (See Fig. A1 for more details).

Fig. A5c shows that 55.5% of user requests fall into the low-creativity category—indicating that most modifications allow little creative input. In this group, the *delete* action predominates, accounting for 51.2% of requests. Conversely, high-creativity requests, which involve more complex transformations, are mainly associated with the *add* action (35.9%). Meanwhile, the medium-creativity category displays a more balanced distribution, with *add* and *delete* actions representing 25.8% and 20.7% of requests, respectively. This distribution suggests that users generally prioritize simple modifications over intricate creative edits.

Table A1. Distribution of user requests and edited images across different data sources in our PSR dataset.

Category	Count	Coverage (%)
<b>Request Statistics</b>		
Total Requests	82,976	100.0
Single-image	71,027	84.4
Multiple-image	11,949	15.6
Historical (Pushshift)	58,624	70.6
Recent Data (2024–2025)	39,610	29.4
<b>Edit Statistics</b>		
Total Edits	305,806	100.0
Historical (Pushshift)	90,466	29.6
Recent Data (2024–2025)	215,340	70.4

Table A2. Distribution of main and subcategories of subjects in image-editing requests.

Category	Count	(%)
<b>People &amp; Related</b>		
Humans & Family	31,914	38.5
Body Parts	9,840	11.9
Clothes & Accessories	2,662	3.2
<b>Text Branding &amp; Abstract</b>		
Text & Logos	8,759	10.6
Special & Misc	2,979	3.6
Abstract & Aesthetic	2,376	2.9
<b>Inanimate Objects</b>		
Tools & Misc	4,021	4.8
Household & Furnishings	3,776	4.6
Vehicles & Transportation	2,728	3.3
<b>Animals</b>		
Pets & Animals	8,411	10.1
<b>Environment &amp; Background</b>		
Natural Environment	3,904	4.7
Lighting & Atmosphere	1,606	1.9
<b>Total</b>	<b>82,976</b>	<b>100.0</b>

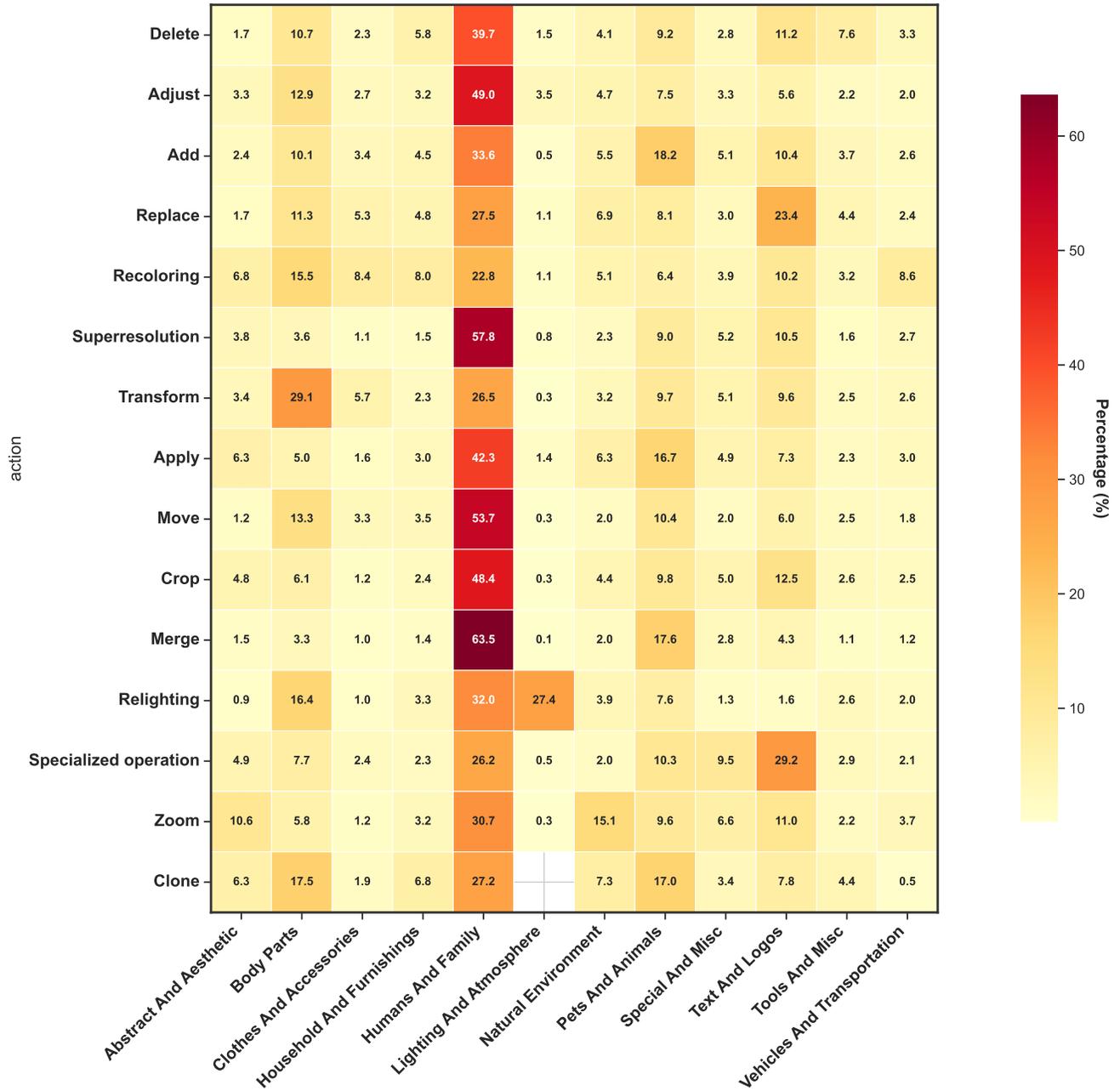


Figure A1. Distribution of Actions Across Different Subject Subcategories



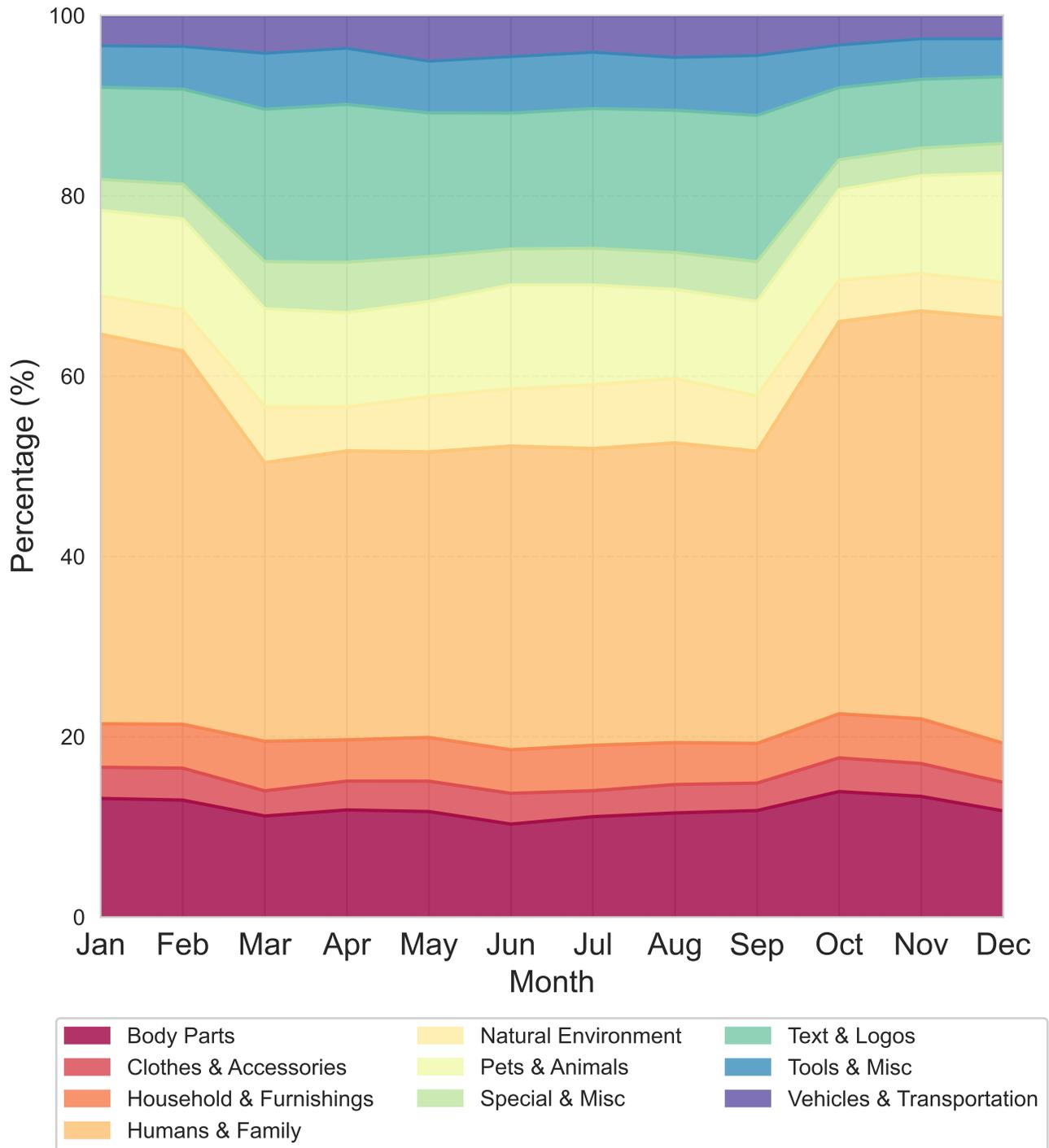
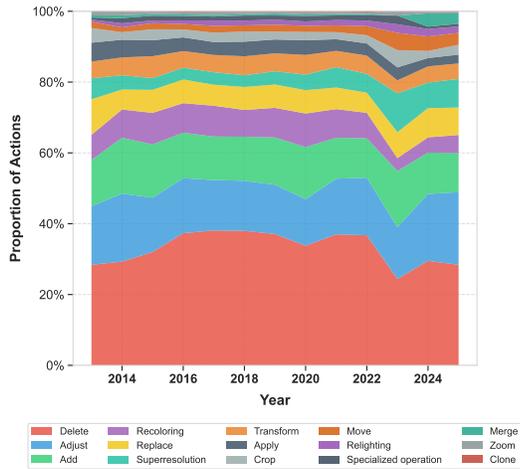


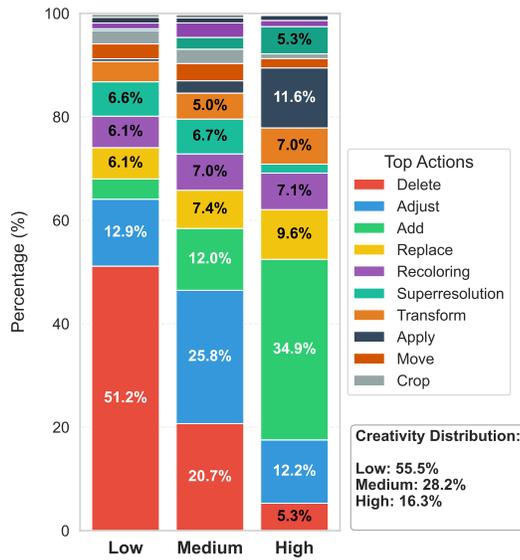
Figure A4. Monthly distribution of subject categories in image editing requests, highlighting trends and variations over time



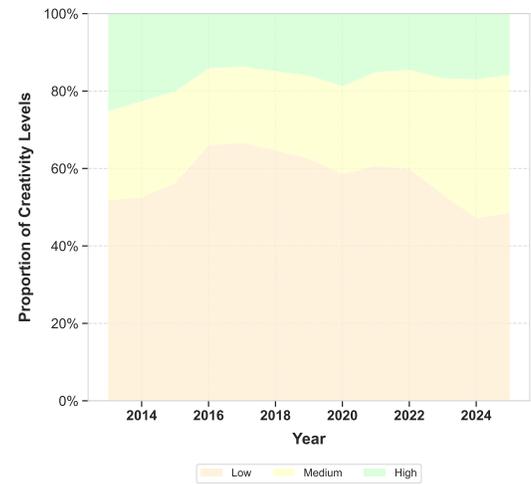
(a) Editing action popularity per year



(b) Most common actions & subjects



(c) Action distribution per creativity level



(d) Creativity level per year

Figure A5. Over 12 years of Reddit data, delete, adjust, and add are the top-3 most wanted actions (a). Specifically, humans, body parts, text, and pets are the most frequent WordNet subjects for such common actions (b). While delete and adjust are the top-2 most common actions in the low - and medium -creativity requests, add takes up the largest share (34.9%) in high -creativity, e.g., inserting some “interesting” background or objects into the scene (c). Most requests (55.5%) require straightforward edits with low creativity (d).

## B. Model Inference Details

In this section, we provide details about the model versions and the temperature.

Table A3. Model Configuration Details

Model	Details
InternVL-2.5-38B	Version: <i>OpenGVLab/InternVL2_5-38B-AWQ</i> Temperature: 0.7 Hosted internally using <a href="#">lmdeploy</a>
GPT-4o-mini	Version: <i>gpt-4o-mini-2024-07-18-global-batch</i> Batch API Temperature: 0 (default) Using the API from Azure OpenAI Service
o1	Version: <i>o1-2024-12-17</i> Temperature: N/A Reasoning Effort: High Using the API from Azure OpenAI Service
Gemini-2.0-Flash-Thinking	Version: <i>gemini-2.0-flash-thinking-exp-01-21</i> Temperature: 0.7 (default) Using the API from Google AI Studio

## C. Prompting Details for Taxonomy Construction

### JSON schema for summarizing an image to JSON (Image-to-JSON)

```
{
  "description": "Brief description of the main content",
  "image_type": "photograph/digital-art/illustration/screenshot/meme",
  "setting": "indoor/outdoor/digital/mixed",
  "location": "beach/office/park/etc",
  "time_of_day": "day/night/unknown",
  "weather": "sunny/cloudy/rainy/not-applicable",
  "has_people": false,
  "people_count": 0,
  "has_adults": false,
  "has_children": false,
  "has_elderly": false,
  "has_groups": false,
  "has_animals": false,
  "has_dogs": false,
  "has_cats": false,
  "has_birds": false,
  "has_wildlife": false,
  "other_animals": [],
  "foreground_objects": [],
  "background_objects": [],
  "prominent_objects": [],
  "dominant_colors": [],
  "lighting": "bright/dim/dark/natural/artificial",
  "has_text": false,
  "text_content": "",
  "text_language": "",
  "mood": [],
  "atmosphere": "",
  "is_nsfw": false,
  "is_violent": false,
  "has_gore": false,
  "has_nudity": false,
  "is_sensitive": false,
  "image_quality": "high/medium/low",
  "orientation": "landscape/portrait/square",
  "tags": [],
  "ai_confidence": "high/medium/low"
}
```

Figure A6. JSON schema for image metadata classification used with InternVL-2.5-38B

## System message for extracting the request's metadata

```
def create_system_message() -> str:
    return """You are an AI system that analyzes image editing requests. Given a textual instruction and an
    image, evaluate the clarity, complexity, and appropriateness of the editing request. Assess the
    instruction's ambiguity (1-5 scale, where 1 is crystal clear and 5 is completely vague), complexity
    level (1-5 scale, where 1 is basic editing and 5 is expert-level), and check for any inappropriate or
    NSFW content. Verify if the image is valid and usable, and determine if the request is actually
    related to image editing. Provide your analysis in the following JSON format, including specific
    reasoning for each field:

    {
      "original_instruction": "Preserved exactly as given to maintain reference point - no modifications or
        interpretations",
      "rewritten_instruction": {
        "text": "Clear, structured version of the original instruction",
        "reasoning": "Clarified version that removes ambiguity, fills in implied steps, and provides specific
          direction. Should be actionable without additional context"
      },
      "missing_details": {
        "items": ["List specific information that would be needed to complete the task but wasn't provided in the
          original instruction"],
        "reasoning": "Identifies gaps that would need to be filled to successfully complete the task"
      },
      "external_references": {
        "value": "True: References external links or comments | False: Self-contained instruction",
        "reasoning": "Identifies if critical information is located outside the main instruction"
      },
      "nsfw_analysis": {
        "value": "True: Contains adult/mature themes | False: Safe for general audience",
        "reasoning": "Evaluates if content contains mature themes, nudity, or adult subject matter"
      },
      "inappropriate_content": {
        "value": "True: Contains harmful/offensive/inappropriate content | False: Appropriate content",
        "reasoning": "Identifies presence of: 1) Harmful content (violence, hate speech, extreme gore) 2)
          Offensive content (discriminatory themes, extreme political content, severe profanity) 3)
          Inappropriate but non-harmful content (crude humor, mild toilet humor, silly/whimsical inappropriate
          gestures, playful trolling). Note: Mild humorous or whimsical content that might be considered 'silly
          inappropriate' (like tongue-in-cheek jokes, mild pranks, or playful memes) should be marked False
          unless they cross into actually offensive territory. Consider context and intent - distinguish between
          harmful inappropriate vs harmless fun"
      },
      "image_editing_relevance": {
        "value": "True: Related to image manipulation | False: Unrelated to image editing",
        "reasoning": "Confirms if the instruction pertains to image editing rather than other topics"
      },
      "image_validity": {
        "value": "True: Image is usable | False: Image is blank/corrupted/missing",
        "reasoning": "Verifies if provided image is suitable for editing"
      }
    }
    """
```

Figure A7. Prompt for extracting basic information from the request

### System message for extracting action verbs from the request

```
def create_system_message(categories_desc) -> str:  
    return f"""Analyze the following image editing instruction and identify which of  
        these specific actions it contains.
```

```
Available categories with examples:  
{categories_desc}
```

#### ANALYSIS GUIDELINES:

1. Evaluate both the original and clarified requests
2. Only include actions that are:
  - Explicitly stated OR
  - Logically necessary to achieve the described result
3. Consider the final image's appearance to identify implicit actions
4. Exclude actions that are:
  - Only potentially useful but not required
  - Vaguely related but not essential
  - Could be used as alternatives

```
Focus only on actual image manipulation actions that match our predefined categories.  
Return your response as a JSON object with an 'actions' array containing only valid  
categories from the list provided.
```

```
"""
```

Figure A8. Prompt for extracting action verbs from an editing request

### System message for assigning a WordNet synset to the subject of an image editing request

```
def create_system_message() -> str:
    return """You are an AI system designed to analyze image editing requests and map
    the subject of the edit to its corresponding WordNet synset. The action
    verb is not mapped-only the subject needs to be processed.

    ### Processing Pipeline:
    1. Candidate Keyword Selection - Extract the most relevant subject from the
    instruction and image. Generate a list of candidate keywords that are highly
    likely to already exist as a WordNet synset. Prioritize concrete nouns and
    commonly recognized entities.
    2. WordNet Synset Matching - Use the refined candidate keywords to query WordNet,
    -confirming and selecting the best-fitting synset(s) for the subject.

    ### Inputs Provided:
    - A textual instruction describing the image edit.
    - An image associated with the request.
    - The previously extracted subject (not the action verb).

    ### Expected Output:
    Generate a structured list of candidate search keywords that:
    - Are highly probable to already exist as a WordNet synset.
    - Accurately represent the subject in the context of the image editing request.
    - Can be directly used for WordNet lookup to retrieve the most relevant synset.

    ### Important Constraints:
    - Do not process or map the action verb-focus solely on the subject.
    - Ensure the keywords are already strong candidates for WordNet synsets before
    attempting lookup.
    - Prefer concrete, commonly used nouns that align well with WordNet's structure.
    """
```

Figure A9. Prompt for assigning WordNet synsets to subjects of editing requests

## Full Details for Subcategories (Part 1)

```
categories_desc = {
  "categories": [
    {
      "category": "Recoloring",
      "definition": "Change the color of an element, object, or text inside the image, but not the whole image",
      "samples": [
        "Can anyone change the dog's fur to black?",
        "Could somebody change the turquoise on the vanity and mirror to white?",
        "Can someone show me how this truck looks in 3 different colors?",
        "Can someone colorize and touch up my grandma?",
        "Can someone please make this grayscale with only the house blue?",
      ],
    },
    {
      "category": "Relighting",
      "definition": "Improve or change the lighting conditions of the scene such as the temperature, color, direction or position of the light source",
      "samples": [
        "pls get rid of the green light or change it to another colour",
        "Can someone relight this photo, removing all harsh shadows",
        "Can someone make lighting better / remove shadows?",
      ],
    },
    {
      "category": "Superresolution",
      "definition": "Modify image so that that the image has a higher resolution and showing clearer, fine details",
      "samples": [
        "How can I increase the pixel count on this picture?",
        "Would a kind soul be able to clean this up with a higher resolution?",
        "Can someone upscale this image to 4K resolution?",
      ],
    },
    {
      "category": "Adjust",
      "definition": "Enhance or correct an entire image's overall appearance by modifying its common properties",
      "samples": [
        "Increase saturation a bit on the elephants",
        "Brighten the shadows by 40% in the portrait",
        "Can someone adjust the lighting/contrast on this?",
      ],
    },
  ],
}
```

Figure A10. Full descriptions for each type of action verb in the data set (Part 1/3).

## Full Details for Subcategories (Part 2)

```
{
  "category": "Delete",
  "definition": "Remove unwanted elements, text, objects, people, or imperfections from the image",
  "samples": [
    "Remove the jacket hanging from the girl's side",
    "Delete the distracting signpost in the background",
    "Please remove the 3rd girl from the left in light blue!",
  ],
},
{
  "category": "Crop",
  "definition": "Trim the edges of an image to make a smaller image to meet specific size requirements",
  "samples": [
    "Crop the photo to eliminate the space to the left and right",
    "Crop to square format for social media",
  ],
},
{
  "category": "Add",
  "definition": "Insert new elements, objects, text, or effects that weren't in the original image",
  "samples": [
    "Insert a ball hitting the tennis racket",
    "Add a copyright watermark to the bottom right",
    "Can someone put a believable tattoo on my daughter?",
  ],
},
{
  "category": "Replace",
  "definition": "Substitute objects or text in the image with something else while keeping the rest of the image unchanged",
  "samples": [
    "Please change the pamphlet she is holding into a dictionary",
    "I hate the background. I would like a neat, white background",
    "Can someone replace the ball with a planet?",
  ],
},
{
  "category": "Apply",
  "definition": "Add filters, styles, or effects that modify the overall appearance of the image",
  "samples": [
    "Add a Gaussian blur to the background",
    "Apply a vintage film effect",
    "Is someone able to help me turn this into a cartoon",
  ],
},
},
```

Figure A10. Full descriptions for each type of action verb in the data set (Part 2/3).

### Full Details for Subcategories (Part 3)

```
{
  "category": "Zoom",
  "definition": "Adjust the image scale to zoom in on a specific area or add new content to mimic a zoom-out action",
  "samples": [
    "Zoom in on the man",
    "Zoom out 50% to show more context",
    "I'm happy to tip if someone is able to zoom this out",
  ],
},
{
  "category": "Transform",
  "definition": "Change the geometric properties (flip, scale, rotate, skew, perspective, distort, warp) of the image or objects",
  "samples": [
    "Flip the photo horizontally",
    "Fix the perspective of the building",
    "Please rotate the box in my hand",
  ],
},
{
  "category": "Move",
  "definition": "Change the position of existing elements within the image while keeping the rest of the image unchanged",

  "samples": [
    "Move the white framed picture to the blue wall",
    "Shift the logo 20 pixels up",
    "Will someone please edit my friend closer to me",
  ],
},
{
  "category": "Clone",
  "definition": "Make more copies of some existing elements inside the image",
  "samples": [
    "Can someone clone my cat",
    "Use a cloning tool to blend grass to cover any patches of dirt",
    "Can someone multiply me and make it look like my arms are interlocked?",
  ],
},
{
  "category": "Merge",
  "definition": "Combine multiple elements or effects from multiple images into a cohesive final image",
  "samples": [
    "Can someone combine these 2 photos?",
    "Please combine so I'm kissing this moose!",
    "Create a panorama from these shots",
  ],
},
{
  "category": "Specialized operation",

  "definition": "Specialized or composite editing operations that don't fit into standard categories",
  "samples": [
    "Can someone vectorize this logo for me without background?",
    "Convert to JPEG format",
    "Can someone make a collage of 12 photos",
  ],
},
]
```

Figure A10. Full descriptions for each type of action verb in the data set (Part 3/3).

## System message for creativity level assignment

```
def create_system_message() -> str:
    return """As an AI assistant, your task is to assign a creativity score to edited
    images based on the diversity of acceptable final versions for a given request
    . The creativity is measured by considering how many different ways the image
    could be edited to fulfill the request.

    - Low Creativity: The request leads to similar edited images with limited
      variations.
    - Medium Creativity: The request allows for some variation but not an extensive
      range.
    - High Creativity: The request can be fulfilled in many different ways leading to
      very different images.

    When evaluating, think about the range of possible acceptable edited images for the
    request.

    Examples:

    1. Request: "Remove the red-eye effect from this photo."
       - The edits will be similar, focusing on correcting the eyes.
       - Creativity Score: Low.

    2. Request: "Transform this portrait into a work of abstract art."
       - There are countless ways to interpret and edit the image.
       - Creativity Score: High.

    3. Request: "Adjust the brightness and contrast to enhance the image."
       - There are some variations in how this can be done.
       - Creativity Score: Medium.

    4. Request: "Crop the image to focus on the main subject."
       - Limited variations in the final image.
       - Creativity Score: Low.

    5. Request: "Add a dramatic sky to this landscape photo."
       - Several ways to interpret a 'dramatic sky.'
       - Creativity Score: Medium.

    6. Request: "Reimagine this landscape in a fantasy setting."
       - Numerous possibilities for how the image could be edited.
       - Creativity Score: High.

    Provide the creativity score (Low, Medium, or High) along with a brief
    explanation for your assessment.
    """
```

Figure A11. Prompt for assigning a creativity level to requests

### System message for comparing different edited images using VLMs

```
def create_system_message() -> str:  
    return """
```

```
You are an image editing evaluation assistant that helps users determine which edited version better fulfills their request. When presented with an original image, an editing request, and two edited versions (A and B), carefully analyze how each edit implements the requested changes. First, examine the specific editing request and how it relates to the original image. Then analyze each edited version, noting strengths, weaknesses, and how closely they match the user's intent. Provide clear reasoning that considers technical quality, aesthetic appeal, and faithfulness to the request. Finally, deliver your verdict in one of three ways: "Image A is better" if the first edit is superior, "Image B is better" if the second edit is superior, or "Tie, both edits are equally good" if they are comparable in quality and adherence to the request.  
"""
```

Figure A12. Prompt for judging different edited images using VLMs.

### Sample Python code for sending multiple images to be evaluated by a judge

```
response = client.chat.completions.create(
    model="MODEL_NAME",
    messages=[
        {"role": "system", "content": system_message},
        {
            "role": "user",
            "content": [
                {
                    "type": "text",
                    "text": f"Analyze the following image editing request and
                        compare the edits:\n\nUser instruction: {
                        original_instruction}"
                },
            ],
        },
        {
            "role": "user",
            "content": [
                {"type": "text", "text": "Source Image"},
                {
                    "type": "image_url",
                    "image_url": {"url": f"{source_image_base64}"},
                },
            ],
        },
        {
            "role": "user",
            "content": [
                {"type": "text", "text": "Edit A"},
                {
                    "type": "image_url",
                    "image_url": {"url": f"{edit1_image_base64}"},
                },
            ],
        },
        {
            "role": "user",
            "content": [
                {"type": "text", "text": "Edit B"},
                {
                    "type": "image_url",
                    "image_url": {"url": f"{edit2_image_base64}"},
                },
            ],
        },
    ],
    tools=[tool],
    tool_choice="auto",
    max_completion_tokens=8192,
)
```

Figure A13. Example code demonstrating how to submit multiple images in one message to a vision-language model (VLM) for judgment in the "VLMs as a Judge" experiment.

## D. Image Generation Details

In this section, we provide details about how we collected the images for all the models.

Table A4. Image Generation Details

<b>Model</b>	<b>Details</b>
GPT-4o	We used ChatGPT’s web interface to generate images.
Gemini-2.0-Flash	We used the official API to generate images.
SeedEdit	We sent images and prompts to SeedEdit’s authors to have images generated locally.
Hugging Face ( 🤖 )	We accessed models through Hugging Face Spaces and inserted the image and prompt (if applicable). Masks and bounding boxes were added manually if needed. See Appendix G.1 for full list of the models.

## E. PSR Action Verbs

Table A5. List of 15 image-editing action verbs from our taxonomy with sample user requests. See Appendix C for the VLM prompts that are used to generate these action labels.

Editing action	Description and Sample Request	Editing action	Description and Sample Request
add	Insert new elements, objects, text, or effects. <i>e.g.</i> “Add a copyright watermark to the bottom right.”	adjust (Fig. 2b)	Modify properties like tones, contrast, and saturation. <i>e.g.</i> “Increase saturation a bit on the elephants.”
apply (Fig. 2f)	Add filters, styles, or effects. <i>e.g.</i> “Apply a vintage film effect.”	clone	Duplicate elements inside the image. <i>e.g.</i> “Use cloning tool to blend grass over dirt patches.”
crop	Trim edges for a smaller image. <i>e.g.</i> “Crop to square format for social media.”	delete (Fig. 2a)	Remove elements, objects, or imperfections. <i>e.g.</i> “Remove the jacket hanging from the girl’s side.”
replace	Substitute objects or text. <i>e.g.</i> “Please change the pamphlet into a dictionary.”	transform	Flip, scale, rotate, or skew elements. <i>e.g.</i> “Fix the perspective of the building.”
move	Reposition elements while keeping the rest unchanged. <i>e.g.</i> “Shift the logo 20 pixels up.”	merge	Combine elements or effects. <i>e.g.</i> “Create a panorama from these shots.”
super-resolution	Increase resolution for clearer details. <i>e.g.</i> “Can someone upscale this image to 4K resolution?”	re-color	Change the color of an element, object, or text. <i>e.g.</i> “Can someone change the dog’s fur to black?”
relight	Adjust lighting conditions. <i>e.g.</i> “Can someone make lighting better / remove shadows?”	zoom (Fig. 2b)	Change scale to focus or zoom out. <i>e.g.</i> “Zoom in on the man.”
specialized	Advanced or composite editing tasks. <i>e.g.</i> “Can someone vectorize this logo without background?”		

## F. Taxonomy Label Failures

We manually evaluated a random subset of 100 samples to assess the accuracy of the WordNet subjects, action verbs, and creativity levels. We observed that  $\approx 5\%$  of the labels generated by GPT-4o-mini contained inaccuracies due to a few harder cases, e.g.:

- **Incorrect use of zoom label.** In the requests *“Help me trick my girlfriend. Hide something funny in this picture that I have to zoom in to show her”* and *“Could someone please photoshop the black bucket in this picture out? You probably have to zoom in to see it,”* an add and delete are needed respectively, not a zoom, despite the word zoom being used in the request.
- **Incorrect use of user WordNet subject.** GPT-4o-mini occasionally labeled user as a subject referring to the Reddit post writer. For instance, in *“Can you take out Buzz and put my sister and me together”*, the correct subject is person (for “my sister and me”) instead of user.

## G. Human Study

In this section, we provide details about the human study. A total of 122 different people participated in our study, from North America, representing two universities and one institution. One-third of the participants are professional image editors and are familiar with image editing techniques.

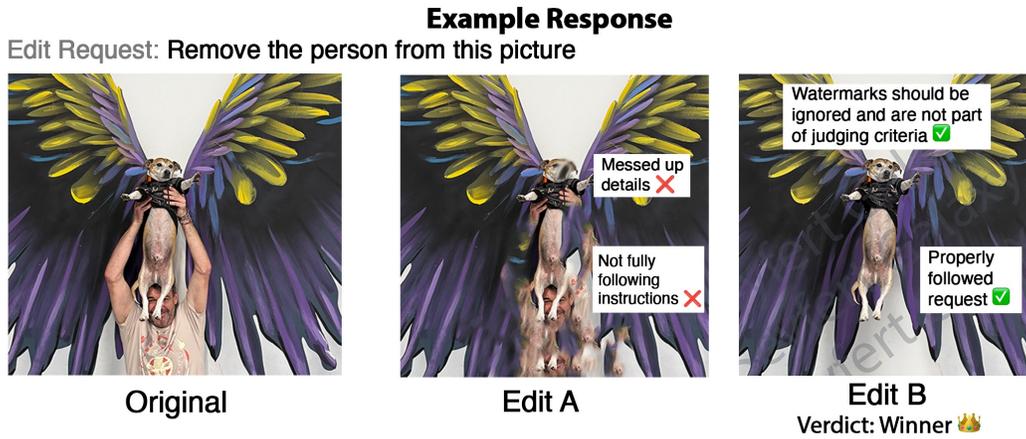


Figure A14. The introduction screen for the human study guides users on how to rate images based on the user's request.

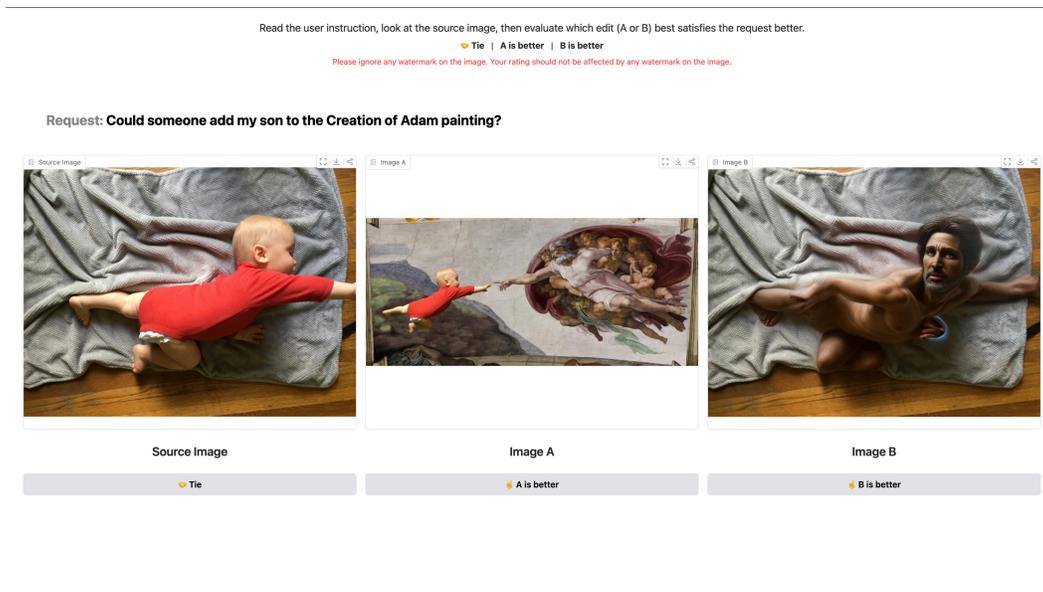


Figure A15. The user interface for the human study displays the original image, the user-provided request, and two edits. Users must decide which edit best satisfies the user's request.

Before starting the survey, users are shown an example of how to judge quality through Fig. A14. Following the survey introduction, the user is shown the original image, the edit request, and the two edited images (Fig. A15). The user then chooses if image A or image B is better, or if they are tied in quality.

## G.1. Details About AI-based Image Editing Tools

Table A6. Number of edits for unique posts generated via different AI tools.

Model	Count	Model	Count
GPT-4o	328	fffiloni/diffusers-image-outpaint	5
Gemini-2.0-Flash	327	ameerazam08/Diffusion-Eraser	4
SeedEdit	274	FLUX.1-Fill-dev	3
CosXL	162	FLUX.1-inpaint-dev	3
InstructPix2Pix	153	diffusers-image-fill	3
MagicQuill	119	IC-Light	3
LEDITS	81	textcutobject	3
FLUX.1-dev-Inpainting-Model-Beta-GPU	52	Finegrain-Object-Cutter	3
remove-photo-object	48	Sketch-Gen	2
BRIA-Eraser-API	47	AnimeGANv2	2
Finegrain-Object-Eraser	45	Image-to-Line-Drawings	2
ReplaceAnything	36	kornia-image-filtering	1
stable-diffusion-xl-inpainting	29	NonLinear-Blurr-Image	1
flux-IP-adapter	27	openfree/ColorRevive	1
CodeFormer	22	fffiloni/InstantIR	1
old_photo_restoration	19	BRIA-Generative-Fill-API	1
flux-fill-outpaint	16	image2coloringbook	1
leditsplusplus	15	yizhangliu/ImgCleaner	1
marcosv/InstructIR	10	remove-photo-object	1
text-guided-image-colorization	10	BRIA-2.3-Inpainting	1
B2BMGMT_Sharpener	10	not-lain/background-removal	1
BRIA-2.2-ControlNet-Recoloring	9	schirrmacher/ormbg	1
foto_filter	7	sketch2lineart	1
turbo_edit	6		

Table A7. Model Performance Comparison (Win Rate as Judged by Humans %)

Model	Total Matches	Human Edit Win Rate	AI Edit Win Rate	Tie Rate
remove-photo-object	3	0.0	100.0	0.0
schirrmacher/ormbg	2	0.0	50.0	50.0
Finegrain-Object-Cutter	1	0.0	100.0	0.0
AnimeGANv2	6	16.7	83.3	0.0
image2coloringbook	5	20.0	20.0	60.0
ameerazam08/Diffusion-Eraser	10	40.0	30.0	30.0
BRIA-Eraser-API	65	43.1	35.4	21.5
openfree/ColorRevive	2	50.0	0.0	50.0
kornia-image-filtering	2	50.0	50.0	0.0
textcutobject	4	50.0	25.0	25.0
IC-Light	4	50.0	50.0	0.0
fffiloni/diffusers-image-outpaint	8	50.0	25.0	25.0
SeedEdit(Simplified Instruction)	443	53.5	39.5	7.0
SeedEdit(Original Instruction)	442	53.6	36.2	10.2
remove-photo-object	60	56.7	21.7	21.7
GPT-4o (Original Instruction)	512	61.1	33.6	5.3
GPT-4o (Simplified Instruction)	502	62.0	32.1	6.0
CodeFormer	46	63.0	26.1	10.9
FLUX.1-dev-Inpainting-Model-Beta-GPU	101	63.4	25.7	10.9
ReplaceAnything	67	65.7	20.9	13.4
turbo_edit	9	66.7	33.3	0.0
BRIA-2.3-Inpainting	3	66.7	33.3	0.0
FLUX.1-Fill-dev	6	66.7	16.7	16.7
Finegrain-Object-Eraser	63	66.7	20.6	12.7
Sketch-Gen	3	66.7	33.3	0.0
Gemini-2.0-Flash (Simplified Instruction)	341	67.2	21.1	11.7
old_photo_restoration	53	67.9	28.3	3.8
MagicQuill	215	68.4	21.9	9.8
BRIA-2.2-ControlNet-Recoloring	13	69.2	15.4	15.4
diffusers-fast-inpaint	17	70.6	23.5	5.9
Image-to-Line-Drawings	7	71.4	14.3	14.3
flux-IP-adapter	43	72.1	20.9	7.0
marcosv/InstructIR	22	72.7	9.1	18.2
Gemini-2.0-Flash (Original Instruction)	340	72.9	20.0	7.1
B2BMGMT_Sharpener	12	75.0	0.0	25.0
flux-fill-outpaint	21	76.2	19.0	4.8
CosXL	334	78.7	14.7	6.6
yizhangliu/ImgCleaner	5	80.0	20.0	0.0
foto_filter	11	81.8	18.2	0.0
text-guided-image-colorization	24	83.3	16.7	0.0
InstructPix2Pix	281	84.0	9.6	6.4
LEDITS	156	84.6	10.3	5.1
leditsplusplus	25	88.0	8.0	4.0
stable-diffusion-xl-inpainting	51	90.2	7.8	2.0
FLUX.1-inpaint-dev	13	92.3	0.0	7.7
fffiloni/InstantIR	1	100.0	0.0	0.0
not-lain/background-removal	1	100.0	0.0	0.0
BRIA-Generative-Fill-API	1	100.0	0.0	0.0
sketch2lineart	1	100.0	0.0	0.0
NonLinear-Blur-Image	2	100.0	0.0	0.0

## G.2. Breakdown of Model Performance by Different Categories

Table A8. Win rate breakdown by different models with at least 30 matchups. SeedEdit leads using Simplified Instructions (SI) with a +3.3% absolute improvement over the Original Instructions (OI).

Model	AI	Human	Tie	Count
SeedEdit (SI) 🏆	39.5	53.5	7.0	443
SeedEdit (OI) 🏆	36.2	53.6	10.2	442
BRIA-Eraser-API	35.4	43.1	21.5	65
GPT-4o (OI)	33.6	61.1	5.3	512
GPT-4o (SI)	32.1	62.0	6.0	502
Old Photo Restoration	28.3	67.9	3.8	53
FLUX.1-Inpainting	25.7	63.4	10.9	101
MagicQuill	21.9	68.4	9.8	215
Gemini 2.0 Flash (OI) ✦	21.1	67.2	11.7	341
Remove-Photo-Object	21.7	56.7	21.7	60
Gemini 2.0 Flash (SI) ✦	20.0	72.9	7.1	340
ReplaceAnything	20.9	65.7	13.4	67
Finegrain-Object-Eraser	20.6	66.7	12.7	63
CosXL	14.7	78.7	6.6	334
LEDITS	10.3	84.6	5.1	156
InstructPix2Pix	9.6	84.0	6.4	281
Stable-Diffusion-XL-Inpainting	7.8	90.2	2.0	51

Table A9. % of votes for human edits and AI edits and for the “Tie” options, categorized by editing actions. “AI Win+Tie” represents the sum of the AI and Tie columns, indicating the percentage of % requests that can already be handled by the 47 AI models.

Action	Human Wins	AI Wins	Tie	AI Win+Tie	no. of votes
<b>Adjust</b>	67.7	25.1	7.2	32.3	1293
<b>Delete</b>	65.1	26.3	8.6	34.9	1123
<b>Add</b>	62.0	30.6	7.4	38.0	1114
<b>Recoloring</b>	68.4	25.2	6.3	31.6	599
<b>Replace</b>	71.2	22.3	6.6	28.8	548
<b>Apply</b>	60.4	30.6	9.0	39.6	399
<b>Transform</b>	70.1	20.8	9.0	29.9	355
<b>Superresolution</b>	70.2	23.6	6.2	29.8	352
<b>Merge</b>	60.7	30.9	8.4	39.3	191
<b>Relighting</b>	66.9	24.2	9.0	33.1	178
<b>Move</b>	68.1	20.2	11.7	31.9	163
<b>Crop</b>	81.4	15.5	3.1	18.6	129
<b>Specialized operation</b>	62.9	26.7	10.3	37.1	116
<b>Zoom</b>	85.3	10.7	4.0	14.7	75
<b>Clone</b>	52.0	28.0	20.0	48.0	25

Table A10. SeedEdit  performance (win rate %), categorized by different action verb types

Action	Human Wins	AI Wins	Tie	AI Win+Tie	Count
<b>Add</b>	48.0	42.0	10.0	52.0	281
<b>Delete</b>	53.5	38.0	8.5	46.5	271
<b>Adjust</b>	54.3	41.1	4.6	45.7	197
<b>Recoloring</b>	63.1	31.1	5.8	36.9	103
<b>Replace</b>	57.0	34.4	8.6	43.0	93
<b>Apply</b>	41.4	54.0	4.6	58.6	87
<b>Transform</b>	62.7	33.9	3.4	37.3	59
<b>Move</b>	68.6	14.3	17.1	31.4	35
<b>Relighting</b>	64.3	28.6	7.1	35.7	28
<b>Merge</b>	37.0	55.6	7.4	63.0	27
<b>Superresolution</b>	52.2	30.4	17.4	47.8	23
<b>Specialized operation</b>	72.7	22.7	4.5	27.3	22
<b>Crop</b>	88.9	5.6	5.6	11.1	18
<b>Zoom</b>	60.0	20.0	20.0	40.0	5
<b>Clone</b>	0.0	100.0	0.0	100.0	1

Table A11. GPT-4o  performance (win rate %), categorized by different action verb types

Action	Human Wins	AI Wins	Tie	AI Win+Tie	Count
<b>Adjust</b>	59.8	35.4	4.8	40.2	336
<b>Delete</b>	68.6	27.7	3.7	31.4	271
<b>Add</b>	50.7	44.3	5.0	49.3	219
<b>Recoloring</b>	63.9	30.4	5.7	36.1	158
<b>Replace</b>	61.5	32.3	6.2	38.5	130
<b>Superresolution</b>	70.3	27.9	1.8	29.7	111
<b>Transform</b>	64.9	28.7	6.4	35.1	94
<b>Apply</b>	53.9	40.8	5.3	46.1	76
<b>Relighting</b>	51.0	37.3	11.8	49.0	51
<b>Merge</b>	54.2	39.6	6.2	45.8	48
<b>Move</b>	51.5	42.4	6.1	48.5	33
<b>Crop</b>	58.6	37.9	3.4	41.4	29
<b>Specialized operation</b>	39.3	42.9	17.9	60.7	28
<b>Zoom</b>	92.0	8.0	0.0	8.0	25
<b>Clone</b>	58.3	33.3	8.3	41.7	12

Table A12. Gemini-2.0-Flash performance (win rate %), categorized by different action verb types

Action	Human Wins	AI Wins	Tie	AI Win+Tie	Count
<b>Adjust</b>	73.6	15.9	10.6	26.4	208
<b>Add</b>	69.0	22.8	8.2	31.0	158
<b>Delete</b>	74.2	21.3	4.5	25.8	155
<b>Recoloring</b>	65.4	24.3	10.3	34.6	107
<b>Replace</b>	67.1	23.5	9.4	32.9	85
<b>Transform</b>	74.6	12.7	12.7	25.4	71
<b>Apply</b>	63.6	21.8	14.5	36.4	55
<b>Superresolution</b>	71.2	21.2	7.7	28.8	52
<b>Merge</b>	71.4	25.7	2.9	28.6	35
<b>Relighting</b>	80.0	11.4	8.6	20.0	35
<b>Move</b>	57.6	24.2	18.2	42.4	33
<b>Crop</b>	75.0	16.7	8.3	25.0	24
<b>Specialized operation</b>	66.7	23.8	9.5	33.3	21
<b>Zoom</b>	83.3	11.1	5.6	16.7	18
<b>Clone</b>	71.4	14.3	14.3	28.6	7

Table A13. Hugging Face performance (win rate %), categorized by different action verb types

Action	Human Wins	AI Wins	Tie	AI Win+Tie	Count
<b>Adjust</b>	75.2	16.5	8.3	24.8	552
<b>Add</b>	73.7	19.7	6.6	26.3	456
<b>Delete</b>	66.9	19.7	13.4	33.1	426
<b>Replace</b>	83.3	11.7	5.0	16.7	240
<b>Recoloring</b>	75.3	19.5	5.2	24.7	231
<b>Apply</b>	71.3	17.7	11.0	28.7	181
<b>Superresolution</b>	72.3	20.5	7.2	27.7	166
<b>Transform</b>	74.8	13.7	11.5	25.2	131
<b>Merge</b>	67.9	19.8	12.3	32.1	81
<b>Relighting</b>	73.4	18.8	7.8	26.6	64
<b>Move</b>	82.3	9.7	8.1	17.7	62
<b>Crop</b>	93.1	6.9	0.0	6.9	58
<b>Specialized operation</b>	71.1	20.0	8.9	28.9	45
<b>Zoom</b>	85.2	11.1	3.7	14.8	27
<b>Clone</b>	20.0	20.0	60.0	80.0	5

Table A14. Human preference win rate (%) by “main category” (all models). AI+Tie denotes share currently handled by AI.

Category	Human Wins	AI Wins	Tie	AI Win+Tie	Count
<b>People And Related</b>	68.5	24.3	7.2	31.5	1992
<b>Inanimate Objects</b>	67.0	23.1	9.8	33.0	864
<b>Animals</b>	59.0	32.9	8.1	41.0	642
<b>Environment And Background</b>	63.5	28.1	8.4	36.5	452
<b>Text Branding And Abstract</b>	65.1	25.2	9.7	34.9	404

Table A15. SeedEdit performance (win rate %), categorized by different subjects

Category	Human Wins	AI Wins	Tie	AI Win+Tie	Count
<b>People And Related</b>	54.2	37.4	8.4	45.8	321
<b>Inanimate Objects</b>	61.7	30.6	7.7	38.3	196
<b>Animals</b>	41.2	49.1	9.7	58.8	165
<b>Environment And Background</b>	51.0	40.2	8.8	49.0	102
<b>Text Branding And Abstract</b>	58.4	32.7	8.9	41.6	101

Table A16. Gemini-2.0-Flash performance (win rate %), categorized by different subjects

Category	Human Wins	AI Wins	Tie	AI Win+Tie	Count
<b>People And Related</b>	72.9	19.2	7.9	27.1	328
<b>Inanimate Objects</b>	62.7	21.2	16.1	37.3	118
<b>Animals</b>	69.7	20.2	10.1	30.3	99
<b>Text Branding And Abstract</b>	72.1	22.1	5.9	27.9	68
<b>Environment And Background</b>	67.2	25.4	7.5	32.8	67

Table A17. GPT-4o performance (win rate %), categorized by different subjects

Category	Human Wins	AI Wins	Tie	AI Win+Tie	Count
<b>People And Related</b>	63.6	31.3	5.1	36.4	514
<b>Inanimate Objects</b>	61.8	31.4	6.8	38.2	191
<b>Animals</b>	53.1	43.4	3.4	46.9	145
<b>Environment And Background</b>	62.7	26.5	10.8	37.3	83
<b>Text Branding And Abstract</b>	61.7	33.3	4.9	38.3	81

Table A18. Hugging Face performance (win rate %), categorized by different subjects

Category	Human Wins	AI Wins	Tie	AI Win+Tie	Count
<b>People And Related</b>	75.4	16.9	7.7	24.6	829
<b>Inanimate Objects</b>	74.1	15.3	10.6	25.9	359
<b>Animals</b>	70.8	20.2	9.0	29.2	233
<b>Environment And Background</b>	69.0	23.5	7.5	31.0	200
<b>Text Branding And Abstract</b>	68.2	17.5	14.3	31.8	154

### G.3. Analysis of Aesthetic Scores

Request	Source	Human Edit	AI Edit	Request	Source	Human Edit	AI Edit
<b>AI Wins with Higher Aesthetic Score</b>							
<i>Place the cat into a space background.</i>				<i>Place my head into a Prussian/Napoleonic military portrait and match the paint texture.</i>			
	Model	SeedEdit			Model	ReplaceAnything	
	Score	4.82	5.60		5.97	Score	4.48
<b>AI Loses despite Higher Aesthetic Score</b>							
<i>Reduce the overexposure in the black and white film image.</i>				<i>Change the color of the sky to purple while keeping the water blue.</i>			
	Model	SeedEdit			Model	MagicQuill	
	Score	3.51	3.90		5.97	Score	5.13
<b>AI Wins with Lower Aesthetic Score</b>							
<i>Remove the background and keyboard from the cat photo to enhance its appearance for framing as a Christmas gift.</i>				<i>Remove the cigarette and smoke from the image.</i>			
	Model	ReplaceAnything			Model	BRIA-Eraser	
	Score	5.27	3.69		5.25	Score	5.55
<b>AI Loses with Lower Aesthetic Score</b>							
<i>Replace the individual in the image with a cartoon character.</i>				<i>Remove the shed and bucket from the background of the photo.</i>			
	Model	CosXL			Model	Remove-Photo-Object	
	Score	4.96	5.60		4.87	Score	5.27

Figure A16. Samples showing that the aesthetic score is not a reliable proxy for human evaluation. The AI edit may win or lose user preference, while the aesthetic score can either increase or decrease relative to the source image.

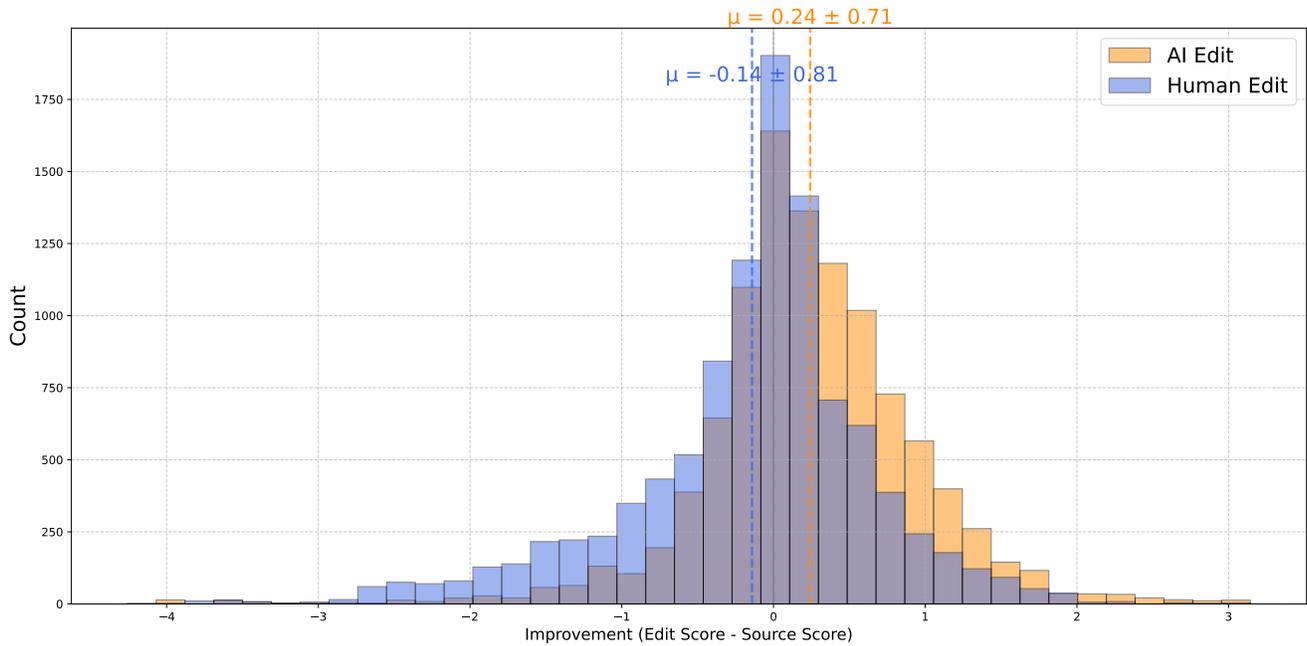


Figure A17. Histogram of Aesthetic Improvement Score (Edited Image Score - Source Image Score)

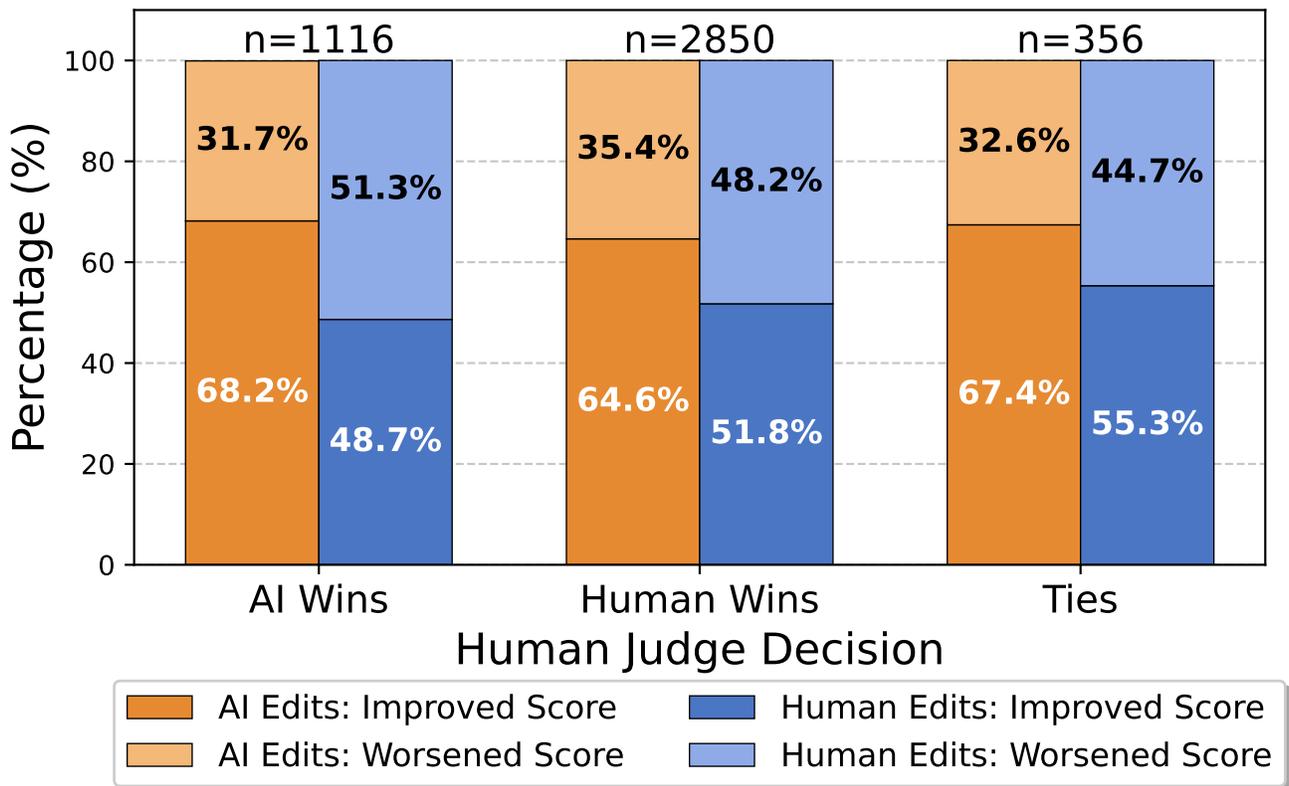


Figure A18. Distribution of aesthetic score changes by different rating outcomes.

## G.4. Additional Results for VLMs-as-a-Judge

Table A19. Cohen’s  $\kappa$  for agreement between human judgment and VLMs, showing generally poor agreement between VLMs and human preferences

Judge	$\kappa$
GPT-4o	0.195 ( $n=4332$ )
o1	0.229 ( $n=4350$ )
Gemini-2.0-Flash	0.177 ( $n=4348$ )

Table A20. Cohen’s  $\kappa$  for agreement between human judgment and VLMs across different model groups. o1 strongly prefers GPT-4o edits, resulting in very poor agreement with humans.

Model	GPT-4o		o1		Gemini-2.0-Flash	
	$\kappa$	$n$	$\kappa$	$n$	$\kappa$	$n$
 SeedEdit	0.226	882	0.287	885	0.203	882
 GPT-4o	0.047	1006	0.054	1013	0.056	1014
 Gemini	0.141	676	0.250	679	0.144	680
 HF	0.203	1768	0.195	1773	0.172	1772

### G.4.1. Sample Reasoning Outputs from VLM Judges

In this section, we provide sample reasoning outputs for each VLM across various agreement settings, compared to human evaluations.

**Can I get the color scheme of this logo changed to shades of gold? - It would be great if I could get it in SVG also. I'll happily do \$20**

Source		<b>Human:</b> Human	<b>o1 Reasoning:</b> The user asked for the original logo to be recolored in gold. Edit A merely provides a color swatch with no actual updated logo. Edit B, despite having an intrusive watermark, does show the logo in a gold color scheme, thus more closely meets the user's request. Overall, B is closer to the requested outcome.
Human Edit		<b>o1:</b> Human	<b>GPT-4o Reasoning:</b> The original image is a logo with the text "That Rhythm Man Guitars" in a stylized font, primarily in shades of teal and cream on a black background. The user requested that the color scheme be changed to shades of gold, and they also asked for the final product in SVG format. Edit A shows a palette of gold shades but does not apply these colors to the logo or provide the logo in any format. This does not fulfill the request, as the color change is not applied to the actual logo. Edit B changes the color scheme of the logo to gold shades, which aligns with the user's request. However, it also includes a watermark "MADE BY ITSYOURBOYANZEY" across the logo, which was not part of the user's request and detracts from the professional appearance of the logo. There is no mention of whether the edit is in SVG format. In terms of fulfilling the color change request, Edit B is closer to what was asked for, despite the unwanted watermark. Edit A does not meet the request at all, as it only shows a gold color palette without applying it to the logo. Therefore, Edit B better fulfills the request as it applies a gold color scheme to the logo, even though it includes an unnecessary watermark.
AI Edit		<b>Gemini-2.0-Flash Thinking:</b> Human	<b>Gemini Reasoning:</b> Edit A provides only a color palette of gold shades, not applying them to the logo itself. It does not fulfill the request to change the logo's color scheme. Edit B successfully changes the logo's color scheme to shades of gold, as requested. While Edit B appears slightly more stylized with a 3D effect than the original, it effectively implements the color change request. Therefore, Edit B is superior because it directly addresses the user's instruction, while Edit A provides an irrelevant color palette.

**Creativity: Medium**

Figure A19. Sample reasoning of VLMs-as-a-judge for the case where all three VLMs agree with the human decision.

**Someone pls make spongbob have a yanke no brim and some other drip**

Source		<b>Human:</b> Human	<b>o1 Reasoning:</b> Both edits add a Yankees cap and extra style, but neither truly removes the brim as requested. Edit A offers more flashy elements like the gold chain and watch, which better satisfies the "some other drip" part of the request. Edit B adds goggles and a stylized background but is otherwise simpler. Overall, Edit A is more in line with the user's desire for additional drip, even though neither edit fully implements the no-brim hat.
Human Edit		<b>o1:</b> AI	<b>GPT-4o Reasoning:</b> The user's request is to edit the image of SpongeBob by giving him a 'Yankee no brim' hat and some 'drip,' which generally refers to fashionable or stylish accessories or clothing. **Original Image Analysis:** The original image shows SpongeBob wearing a 'Supreme' hoodie, which is already a brand associated with fashionable streetwear. However, the hat and additional accessories requested by the user are not present. **Edit A Analysis:** - The image features a hat with the New York Yankees logo, which aligns with the request for a Yankee hat. However, it still has a brim, which slightly deviates from the 'no brim' specification. - Additional 'drip' includes red-tinted sunglasses, adding a trendy style element. - The background is changed to a crumpled paper texture, which adds an aesthetic appeal but is not part of the request. **Edit B Analysis:** - The image also features a New York Yankees hat, which fits the user's request. Like in Edit A, the hat has a brim, which again does not fully meet the 'no brim' specification. - Additional 'drip' includes gold chains and bracelets, which enhance the fashionable style significantly more than Edit A. - The image retains the plain background similar to the source, focusing more on the style elements added to SpongeBob rather than the surroundings. **Verdict:** While both edits feature the Yankee hat with a brim, Edit B excels in adding more 'drip,' such as multiple gold accessories, which better aligns with the user's request for stylish enhancements. Therefore, Image B is better.
AI Edit		<b>Gemini-2.0-Flash Thinking:</b> Human	<b>Gemini Reasoning:</b> The user requested to add a 'yanke no brim' (Yankees no brim hat) and 'some other drip' to the original Spongebob image. Both edits successfully added a Yankees no-brim hat. **Edit A:** Added a navy blue Yankees no-brim hat, a gold chain necklace with a large pendant, gold bracelets on both wrists, and a gold watch on the right wrist. This edit interprets 'drip' as expensive-looking jewelry, enhancing the character's 'drip' through accessories associated with wealth and status. The additions are stylistically consistent with the original cartoon. **Edit B:** Added a black Yankees no-brim hat and a red 'Supreme' branded headband worn above the hat, and also stylistically fits the cartoon image. While both edits fulfill the request by adding a hat and 'drip,' Edit B is slightly more creative and visually impactful. The 'Supreme' headband is a more distinct and contemporary representation of 'drip' in streetwear fashion compared to the more generic gold jewelry in Edit A. The headband also adds a more playful and slightly absurd element, which aligns well with the original image's humorous context.

**Creativity: High**

Figure A20. Sample reasoning of VLMs-as-a-judge for the case with mixed results, where some of the VLMs agree with the human decision.

(venmo \$5) could you take away the green cast and overexposure of this pic of my dog as a puppy?

Source		<p><b>Human:</b> Human</p> <p><b>o1:</b> AI</p> <p><b>GPT-4o:</b> AI</p> <p><b>Gemini-2.0-Flash Thinking:</b> AI</p>	<p><b>o1 Reasoning:</b> Both edits remove the worst of the greenish tint, but Edit B does a slightly better job bringing back natural black-and-tan coloring. In Edit A, the fur has some pinkish patches and the background looks flatter. Edit B, although not perfect, is more balanced overall and closer to realistic tones, so it better fulfills the request to eliminate the green cast and fix the overexposure.</p> <p><b>GPT-4o Reasoning:</b> The original image has a noticeable green color cast and appears overexposed, particularly affecting the dog's fur, which appears unnatural, and the overall brightness of the image. Edit A: This version reduces the green color cast substantially, making the colors appear more neutral, which helps in making the dog's fur look more natural. The image appears less overexposed, with a more balanced contrast, but there is a watermark that says "Sample" over the image, which detracts from its final quality. Edit B: This version also removes the green cast effectively and adjusts the exposure, resulting in a more natural look for the dog's fur and the surrounding area. There is no watermark, giving it a cleaner appearance than Edit A. However, the image still has some color inconsistencies, especially around the dog's fur which appears slightly orange in some areas, which might not be accurate. Overall, both edits significantly improve on the original by addressing the green cast and overexposure. However, Edit B offers a cleaner final product without the distraction of a watermark, making it a more complete and polished edit despite minor issues with color accuracy.</p> <p><b>Gemini Reasoning:</b> The user requested removal of a green cast and overexposure from a photo of a puppy. Original image exhibits a strong green color cast and highlights appear blown out, indicating overexposure. Edit A successfully removes the green cast and reduces the overexposure, resulting in a more natural color balance and better contrast. The puppy and the background elements are now more clearly visible with improved detail. Edit B converts the image to black and white. While this eliminates the green cast and addresses overexposure by simplifying the tonal range, it fundamentally changes the image's aesthetic and removes color information. Edit A is superior because it directly addresses the user's request by correcting the color and exposure issues while maintaining the original color context of the photograph, which is generally preferred when asking to 'fix' color casts. Edit B is a less direct interpretation of the request and makes a more drastic stylistic change.</p>
Human Edit			
AI Edit			<p><b>Creativity:</b> Low</p>

Figure A21. Sample reasoning of VLMs-as-a-judge for the case where all three VLMs disagree with the human decision.

idk if this does belong here, but I did a photo of some stars and I wanted to edit it on lightoom to remove the orange on top and make the stars more visible. thanks

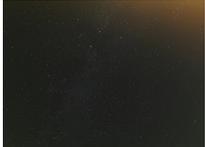
Source		<p><b>Human:</b> Human</p> <p><b>o1:</b> Human</p> <p><b>GPT-4o:</b> Human</p> <p><b>Gemini-2.0-Flash Thinking:</b> AI</p>	<p><b>o1 Reasoning:</b> Both edits attempt to eliminate the orange glow and bring out the stars, but Edit B does so more completely and evenly. In Edit A, the orange band is still quite prominent and a dark patch in the center detracts from overall clarity, while the stars themselves do not pop as vividly. Edit B does a better job of removing the orange cast and boosting star contrast and visibility, producing a cleaner, more uniform night sky.</p> <p><b>GPT-4o Reasoning:</b> The original image depicts a night sky with stars scattered across a black background. However, it has a noticeable orange tint at the top, likely due to light pollution or lens flare, which detracts from the clarity and visibility of the stars. <b>User's request:</b> The user wants to remove the orange tint at the top and make the stars more visible, likely seeking a clearer, more natural representation of the night sky. <b>Analysis of Edit A:</b> <b>Orange Tint Removal:</b> The orange tint is still visible in Edit A, albeit slightly reduced. The top of the image still has a noticeable orange hue which suggests the edit did not fully address the user's request to remove it. <b>Star Visibility:</b> The stars in Edit A appear slightly enhanced compared to the original. However, the visibility improvement is minimal, and the stars do not pop significantly more than in the source image. <b>Analysis of Edit B:</b> <b>Orange Tint Removal:</b> Edit B successfully removes the orange tint from the top, resulting in a more uniform dark sky. This aligns well with the user's request. <b>Star Visibility:</b> The stars are much more visible in Edit B, with greater contrast against the dark sky. The image has a clearer and more dynamic look, making the stars stand out more prominently. <b>Conclusion:</b> Edit B more effectively fulfills the user's request by removing the orange tint completely and enhancing the visibility of the stars significantly more than Edit A.</p> <p><b>Gemini Reasoning:</b> The user requested to remove the orange tint at the top of the star photo and make the stars more visible. Both edits successfully removed the orange tint. Edit A shifted the color balance towards a strong blue hue across the entire image, which while removing the orange, introduces an unnatural blue tint to the night sky. This edit also darkens the image significantly, making the background almost completely black, which may lose some subtle details of the night sky. Edit B, on the other hand, effectively removes the orange tint and increases the visibility of stars by enhancing their brightness without introducing a strong color cast. Edit B maintains a more natural dark tone for the night sky, making the stars stand out more subtly and realistically. Therefore, Edit B is better because it achieves the user's goals while preserving a more natural and aesthetically pleasing look for a night sky photo compared to the overly blue and dark Edit A.</p>
Human Edit			
AI Edit			<p><b>Creativity:</b> Low</p>

Figure A22. Sample reasoning of VLMs-as-a-judge for the case with mixed results, where some of the VLMs agree with the human decision.

Can someone Jurassic-Park this pic of my kids in the jungle? Add a dinosaur or something? Thx.

Source		<p><b>Human:</b> AI</p>	<p><b>o1 Reasoning:</b> Both edits place dinosaurs into the original jungle scene, but Edit B creates a bigger, more dramatic 'Jurassic Park' effect, complete with realistic color-matching, large-scale dinosaurs, and a cinematic feel. Edit A has a single smaller dinosaur that looks somewhat less integrated and less imposing, while Edit B's two dinosaurs match the scale and atmosphere of the movie style more closely. Overall, Edit B better fulfills the request to 'Jurassic-Park' the image.</p>
Human Edit		<p><b>o1:</b> AI</p> <p><b>GPT-4o:</b> AI</p> <p><b>Gemini-2.0-Flash Thinking:</b> AI</p>	<p><b>GPT-4o Reasoning:</b> The user's request is to "Jurassic-Park" the picture by adding a dinosaur or something similar, enhancing the jungle setting with a prehistoric theme. <b>**Edit A Analysis:**</b> - Two large, realistic dinosaurs have been added, one prominently in the foreground and another in the mid-ground. - The dinosaurs are integrated well into the scene with proper scaling and shadowing, making them appear part of the environment. - The dinosaurs are detailed and match the lighting and color palette of the original image, creating a dramatic and convincing prehistoric atmosphere. <b>**Edit B Analysis:**</b> - A smaller dinosaur has been added in the mid-ground towards the right side of the image. - The dinosaur is less prominent and not as integrated into the scene as in Edit A. It lacks the dramatic impact and presence that Edit A's dinosaurs have. - The dinosaur in Edit B is less detailed, and the lighting doesn't match the foggy, mystical atmosphere of the jungle as effectively. <b>**Comparison and Verdict:**</b> - Edit A provides a more immersive and visually striking transformation, with the dinosaurs appearing both realistic and appropriate for the jungle setting. - Edit B falls short in terms of visual integration and dramatic effect, as the dinosaur appears more like an afterthought rather than a central feature. Overall, Edit A better fulfills the user's request by providing a more engaging and visually cohesive Jurassic theme.</p>
AI Edit		<p><b>Gemini Reasoning:</b> Both edits successfully incorporate dinosaurs into the jungle scene as requested, fulfilling the user's 'Jurassic Park' theme. However, Edit B is superior due to several factors. Edit B includes two dinosaurs, a large one prominently placed near the children and a second one in the background, enhancing the 'Jurassic Park' feel by suggesting a more populated and potentially dangerous environment. The dinosaurs in Edit B are also better integrated in terms of lighting and perspective, making them look more natural within the scene compared to Edit A. In Edit A, the single dinosaur feels somewhat isolated and its integration is less seamless. The presence of two dinosaurs in Edit B creates a more dynamic and engaging composition that aligns more closely with the adventurous and slightly menacing atmosphere associated with Jurassic Park, making it a more effective and impactful edit overall.</p>	<p><b>Creativity:</b> High</p>

Figure A23. Sample reasoning of VLMs-as-a-judge for the case where all three VLMs agree with the human decision.

My girlfriend needs to be cheered up! Please have some fun with this picture of our cat!

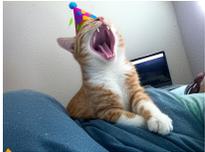
Source		<p><b>Human:</b> Human</p>	<p><b>o1 Reasoning:</b> Edit A playfully adds a party hat, keeping the main focus on the cat's open-mouthed pose. Edit B takes a bolder comedic approach by placing the cat in a humorous new setting that feels more imaginative and over-the-top, which aligns well with the goal of cheering someone up. Both edits capture a fun vibe, but the creative and amusing scenario in Edit B likely inspires more laughter and excitement.</p>
Human Edit		<p><b>o1:</b> Human</p> <p><b>GPT-4o:</b> Human</p> <p><b>Gemini-2.0-Flash Thinking:</b> Human</p>	<p><b>GPT-4o Reasoning:</b> The user requested an edit to "have some fun" with the picture of their cat, presumably to cheer someone up. This request suggests a playful, humorous, or creative edit that would bring joy or laughter. <b>Edit A Analysis:</b> - In Edit A, a colorful party hat has been added to the cat's head. This addition is simple yet effective for creating a playful and cheerful atmosphere. The party hat is well-integrated and fits naturally on the cat's head, enhancing the theme of fun with minimal effort. The technical quality of the edit is good, and it maintains the original image's clarity and focus on the cat. <b>Edit B Analysis:</b> - Edit B places the cat among a group of lions, which is a creative and humorous juxtaposition. The cat's head is seamlessly blended with the lions, creating an amusing scene where the cat appears to be part of a pride. This edit is more elaborate and imaginative, likely to provoke laughter due to the absurdity of a domestic cat mingling with wild lions. The technical execution is well done, and the absurdity aligns well with the request for a fun edit. <b>Comparison:</b> - Edit A is straightforward, focusing on a lighthearted addition to the original image by using a party hat, which is a classic symbol of celebration and fun. - Edit B goes a step further by creating an unexpected and humorous scenario. It involves more creativity and could be more effective in cheering someone up through its imaginative approach and the visual comedy of the cat being among lions. Both edits are technically well-executed, but Edit B offers a more engaging and humorous interpretation of the request, likely to have a greater impact in terms of cheering someone up.</p>
AI Edit		<p><b>Gemini Reasoning:</b> The user requested to 'have some fun' with a picture of their cat to cheer up their girlfriend. Edit A adds a party hat to the cat, which is a simple and cheerful edit, suggesting a celebratory or fun occasion. Edit B places the cat within a group of lions in a wild setting. This edit is more creative and humorous due to the unexpected and absurd juxtaposition of a domestic cat among lions. The cat's yawning expression in the original image can be interpreted as a roar in this new context, adding to the humor. While both edits attempt to be fun, Edit B is more imaginative, unexpected, and likely to elicit a stronger positive emotional response due to its comedic nature. Therefore, Edit B better fulfills the request to 'have fun' with the picture and cheer someone up by being more creative and humorous.</p>	<p><b>Creativity:</b> High</p>

Figure A24. Sample reasoning of VLMs-as-a-judge for the case where all three VLMs agree with the human decision.

**My girlfriend needs to be cheered up! Please have some fun with this picture of our cat!**



**Human:**  
Human

**o1:**  
Human

**GPT-4o:**  
Human

**Gemini-2.0-Flash Thinking:**  
Human



**o1 Reasoning:**

The user wants a fun, cheerful edit. Edit A simply outlines the cat in pink, which doesn't add much humor. Edit B places the cat among a group of lions, matching the spirit of the request by creating a playful, humorous scene more likely to cheer someone up.

**GPT-4o Reasoning:**

The editing request was to have fun with a picture of a cat to cheer up the user's girlfriend. This suggests a playful or humorous alteration to the image. **\*\*Edit A Analysis:\*\*** Edit A features a simple and minimal change to the original image. The cat is outlined with a pink border, which doesn't contribute significantly to a humorous or fun effect. While it adds a pop of color, it lacks creativity or an element that would likely elicit a strong emotional response such as laughter or surprise. **\*\*Edit B Analysis:\*\*** Edit B places the cat among a group of lions, creating a humorous and unexpected scene. The positioning of the cat, who is yawning or perhaps meowing, among the lions adds a humorous juxtaposition and plays on the idea of the cat being part of a pride of lions. This edit is more imaginative and likely to make someone smile or laugh due to the absurdity and creativity of the composition. **\*\*Conclusion:\*\*** Edit B is more in line with the request to "have fun" with the picture. It uses creativity and humor effectively by placing the cat in an unexpected and funny situation. Therefore, Edit B better fulfills the user's request to cheer up the girlfriend by having fun with the image.

**Gemini Reasoning:**

The user's request was to 'have some fun' with the picture of their cat to cheer up their girlfriend. Edit A simply outlines the cat with a pink border, which does not introduce any humor or fun element and doesn't significantly alter the original image in a playful way. Edit B, however, cleverly inserts the cat into a scene with lions, creating a humorous and unexpected scenario. The juxtaposition of a domestic cat among wild lions is inherently funny and directly addresses the request to 'have fun' with the picture. Edit B is much more likely to cheer someone up due to its creative and amusing nature, whereas Edit A is a very minimal and uninspired edit.

Figure A25. Sample reasoning of VLMs-as-a-judge for the case where all three VLMs agree with the human decision.

**Hey can you guys make this boring beach photo I took look epic.**



**Human:**  
AI

**o1:**  
Human

**GPT-4o:**  
AI

**Gemini-2.0-Flash Thinking:**  
AI



**o1 Reasoning:**

Both edits offer a more dramatic look than the original black-and-white shot. However, Edit A adds a beautiful warm-to-cool sky gradient while keeping the fisherman silhouette and the overall composition of the original, so it remains recognizable as the user's own photo while making it feel more vivid. By contrast, Edit B is extremely dramatic but replaces almost everything—there is now a lighthouse, crashing surf, and a completely different sky. Although it looks highly 'epic,' it no longer resembles the original photo or preserves the fisherman. For simply 'making the user's photo look epic' without departing from its core, Edit A better fulfills the request.

**GPT-4o Reasoning:**

The user's request was to transform a "boring beach photo" into something "epic." The original image is a monochrome picture of a beach with a lone person standing far away and a couple of birds in the sky, which indeed feels a bit plain due to lack of color and dramatic elements. Edit A introduces a vibrant and dynamic sunset scene with bold colors ranging from deep blues and purples to fiery oranges and pinks. It also adds a lighthouse, large crashing waves, and numerous birds in the sky, creating a dramatic and lively atmosphere. This edit significantly enhances the image's visual appeal by adding multiple elements that evoke a sense of wonder and excitement. The technical quality of the colors and clarity is high, and the scene looks quite vibrant and lively. Edit B maintains the original composition but adds a colorful gradient sky with shades of orange and blue, suggesting a sunrise or sunset. This edit keeps the original beach silhouette and the person with the fishing rod, adding a warm and inviting feel without altering the image's content too much. The change in sky color makes the scene more engaging than the original while preserving its calm and serene atmosphere. While both edits improve on the original, Edit A takes more liberties and adds more elements to create an "epic" scene, fitting the user's request more effectively by transforming the entire setting into a more dramatic and engaging one. Edit B, while aesthetically pleasing, remains closer to the original and relies solely on color enhancement to achieve the change, which may not be as impactful in terms of achieving an "epic" transformation.

**Gemini Reasoning:**

The user requested to make their 'boring beach photo' look 'epic'. The original image is a monochromatic, minimalist beach scene with a lone fisherman. Edit A introduces a vibrant sunset color palette, changing the sky to a gradient of blue, orange, and yellow. This adds warmth and interest but retains the simple composition of the original. Edit B also uses a sunset color palette, but it's much more dramatic, featuring deep purples, oranges, and yellows in dynamic clouds. Furthermore, Edit B significantly alters the scene by adding a lighthouse, larger crashing waves, and numerous birds, creating a sense of grand scale and drama. While Edit A improves the image by adding color, Edit B more effectively fulfills the request to make the photo 'epic' by introducing more dramatic elements and a more impactful composition that conveys a sense of awe and grandeur. The addition of the lighthouse and powerful waves in Edit B contribute significantly to the 'epic' feeling, surpassing the more subtle changes in Edit A.

**Creativity:** High

Figure A26. Sample reasoning of VLMs-as-a-judge for the case with mixed results, where some of the VLMs agree with the human decision.

## G.4.2. VLM Judges Fail to Detect Changes and Alterations in Subject Identity

**Can you guys make this picture more interesting? Go crazy!**

Source		<b>Human:</b> Human	<p><b>o1 Reasoning:</b> Both edits take the request to “go crazy” in different directions. Edit A transforms the entire setting into a vibrant nighttime scene with fireworks and colorful lighting, giving it a fun, festival-like energy that highlights the action. Edit B introduces a charging rhinoceros into the yard, which is certainly an eye-catching and humorous scenario. While both approaches are creative, Edit A pushes the setting further with dramatic color and atmosphere, making the image feel especially lively and over-the-top. Overall, Edit A delivers a bigger transformation and stands out as the more striking, celebratory “crazy” edit.</p> <p><b>GPT-4o Reasoning:</b> Edit A has transformed the image into a vibrant and lively scene with fireworks in the sky, neon lighting on the buildings, and a cityscape in the background. This edit significantly alters the original suburban setting into a more festive and urban environment, making the picture appear more dynamic and exciting. Edit B introduces a large rhinoceros charging across the lawn, which adds an element of surprise and humor to the image. The addition of the rhinoceros is unexpected and certainly makes the image more interesting by creating a sense of action and urgency. Both edits fulfill the request to make the picture more interesting. Edit A does so by adding festive elements and altering the background to create a more celebratory atmosphere. Edit B achieves interest by introducing an unusual and exciting element into the setting. In terms of creativity and making the image “more interesting,” Edit A might appeal to those who enjoy vibrant and colorful edits, while Edit B might appeal to those who appreciate humor and the unexpected. Both edits are technically well-executed, but the choice between them depends on the user’s taste in what constitutes “interesting.” Given the request to “go crazy,” Edit A seems to push the boundaries more by completely transforming the scene, whereas Edit B is more subtle in its craziness with the single addition of a rhinoceros. Verdict: Image A is better as it more drastically transforms the scene and aligns with the user’s instruction to “go crazy.”</p> <p><b>Gemini Reasoning:</b> Edit A adds fireworks and a city skyline at night, transforming the daytime suburban scene into a festive nighttime celebration. This is visually appealing and fulfills the “go crazy!” request by adding dramatic and exciting elements. Edit B inserts a rhinoceros into the scene, creating a surreal and humorous image. Both edits successfully make the image more interesting as requested. However, Edit A’s changes are more visually coherent and aesthetically pleasing, creating a more impactful and festive “crazy” atmosphere compared to the absurdity of Edit B.</p>
Human Edit		<b>GPT-4o:</b> AI	
AI Edit		<b>Gemini-2.0-Flash Thinking:</b> AI	

Creativity: High

Figure A27. Sample reasoning by VLMs-as-a-judge for cases where all VLMs disagree with human judgments, primarily due to VLMs ignoring subtle differences between images.

**Can anyone please add a morning sky to my image?**

Source		<b>Human:</b> Human	<p><b>GPT-4o Reasoning:</b> The original image depicts a hiker looking out over a mountainous landscape with a blank white sky. The request is to add a morning sky to enhance the scene and make it more visually appealing. Edit A introduces a dramatic sky with deep orange and yellow hues, suggesting a sunrise or early morning sky. The colors are vibrant and the clouds are well-defined, creating a striking contrast with the landscape. However, the overall lighting of the scene seems slightly dark, especially in the foreground, which might not entirely match the brightness of a typical morning. Edit B presents a vibrant sky as well, with shades of pink, purple, and orange. The clouds are beautifully textured and realistic, giving the impression of a colorful morning sky. Additionally, the lighting is adjusted so that the foreground is brighter, matching the morning theme more accurately. This edit also maintains a natural look in the landscape, making it feel more consistent with the time of day depicted. While both edits successfully add a morning sky, Edit B is superior due to its more natural integration of lighting and color, which better fulfills the request for a morning atmosphere.</p> <p><b>Gemini Reasoning:</b> The user requested to add a 'morning sky' to the image. The original image has a washed-out sky. Both edits replaced the sky with a colorful sunrise/sunset which is a reasonable interpretation of a 'morning sky'. Edit A adds a golden and orange sky but shows some noticeable haloing around the mountain edges and the foreground remains quite dark, not matching the intense sky light. Edit B adds a vibrant red and orange sky with detailed clouds. The integration of the sky in Edit B appears more seamless with the mountains, and the overall image has slightly better contrast and detail in the foreground, although the lighting still doesn't fully match the dramatic sky. Edit B also seems to have slightly altered the composition, showing more of the path on the right, which wasn't requested, but the primary request was about the sky replacement. Overall, Edit B provides a more visually appealing and better-integrated sky despite the minor compositional change and lighting mismatch.</p>
Human Edit		<b>GPT-4o:</b> AI	
AI Edit		<b>Gemini-2.0-Flash Thinking:</b> AI	

Creativity: High

Figure A28. Sample reasoning by VLMs-as-a-judge for cases where all VLMs disagree with human judgments, primarily due to VLMs ignoring subtle differences between images.

**Can someone please make the rings look more brilliant/shiny? TIA**



**Human:**  
Human

**o1:**  
AI

**GPT-4o:**  
AI

**Gemini-2.0-Flash Thinking:**  
AI

**o1 Reasoning:**

In Edit A, the ring has been noticeably brightened and polished—its gold color is enhanced, contrast and clarity have been boosted, and the diamond now sparkles with highlights. It looks like a crisp jewelry advertisement and very clearly fulfills the user’s desire for it to look more brilliant and shiny. Edit B, by comparison, shows only a subtle increase in saturation and shine and is less striking overall. Since the user specifically asked for the ring to be more brilliant/shiny, Edit A captures that request far more effectively.

**GPT-4o Reasoning:**

The user requested an enhancement to make the rings look more brilliant or shiny. Looking at the original image, the rings appear a bit dull due to the lighting and focus, with the background being more prominent. Edit A Analysis: Edit A has increased the brightness and saturation, making the yellow gold band more vivid. However, the diamond and overall shine of the rings still appear somewhat subdued. The edit has enhanced the color but not significantly increased the brilliance or shine of the rings. The sand and overall background remain largely unchanged, which does not contribute to the rings’ brilliance. Edit B Analysis: Edit B has significantly enhanced the brilliance of the rings. The diamond exhibits a sparkle effect, reflecting light, and the gold band is much shinier compared to the original. The lighting seems adjusted to focus more on the rings, drawing attention to the requested shine. The sand appears more sparkly, contributing to the overall aesthetic of brilliance. This edit fulfills the request for making the rings look shiny and brilliant more effectively. Verdict: Edit B is better as it more thoroughly fulfills the user’s request by enhancing both the diamond and the gold band to appear more shiny and brilliant.

**Gemini Reasoning:**

The user requested that the rings look more brilliant and shiny. Both edits attempted to fulfill this request by brightening the rings. Edit A went further by adding significant highlights and starburst effects to the diamond and surrounding stones, which effectively conveys the brilliance and shininess of jewelry. Edit B made the rings brighter, but lacked the specific sparkling highlights that make diamonds appear brilliant. Therefore, Edit A is more successful in achieving the desired effect.

**Creativity:** Medium

Figure A29. Sample reasoning by VLMs-as-a-judge for cases where all VLMs disagree with human judgments, primarily due to VLMs ignoring subtle differences between images.

**Can someone please remove the glare of the window on this picture? Thanks!**



**Human:**  
Human

**o1:**  
AI

**GPT-4o:**  
AI

**Gemini-2.0-Flash Thinking:**  
AI

**o1 Reasoning:**

Both edits address the user’s request to remove the bright outlines and shapes at the top of the picture. Edit A appears to do a minimal retouch to remove the reflections, preserving the original contrast levels but leaving a slightly uneven sky gradient. Edit B removes all traces of glare and creates a very smooth gradient, while also sharpened or increased contrast in the mountains, resulting in a more dramatic final image. From a purely technical perspective, Edit B more thoroughly eliminates the window glare without artifacts and yields a cleaner, more polished sky area. Therefore, B is the stronger edit.

**GPT-4o Reasoning:**

The user has requested to remove the glare caused by the window, which is visible in the original image as a bright reflection in the sky area. This glare distracts from the view of the mountains below. Edit A: In this version, the glare has been reduced significantly. The bright spots in the sky area have been blended with the surrounding colors, making the glare less noticeable. However, there is still a slight difference in texture where the glare was originally located, which might be noticeable upon close inspection. Edit B: This version has completely removed the window glare, leaving a smooth gradient sky without any noticeable brightness differences. The overall image appears more natural and cohesive, with no visible remnants of glare. The edit maintains the integrity of the mountain view, enhancing the aesthetic appeal of the image. Both versions effectively address the user’s request to remove the window glare, but Edit B does so with a higher degree of technical precision and aesthetic quality compared to Edit A.

**Gemini Reasoning:**

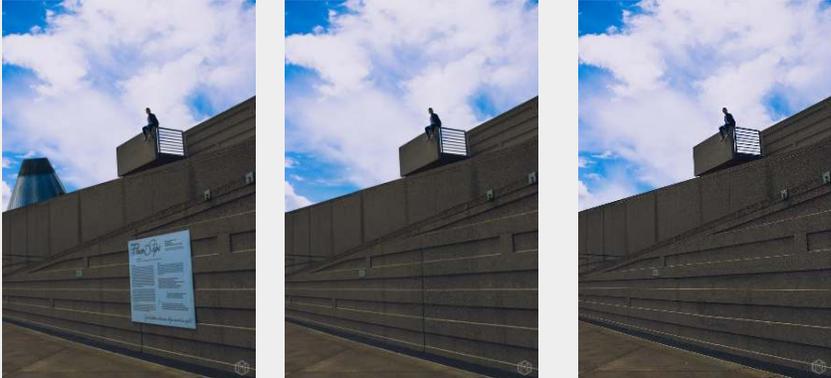
The user requested the removal of window glare from the image. The original image displays significant glare artifacts in the upper portion. Edit A successfully removes the glare but appears to have applied a gradient or perhaps replaced the sky, resulting in a less natural look in the upper sky area. Edit B also effectively removes the glare but preserves the natural appearance of the sky gradient and the overall lighting of the scene much better than Edit A. Edit B maintains more detail and a more realistic tone.

**Creativity:** Low

Figure A30. Sample reasoning by VLMs-as-a-judge for cases where all VLMs disagree with human judgments, primarily due to VLMs ignoring subtle differences between images.

### G.4.3. Hallucinations by VLMs Acting as Judges

o1 VLM-as-a-Judge Hallucinations



**Source**                      **Edit A (Human)**                      **Edit B (AI)**

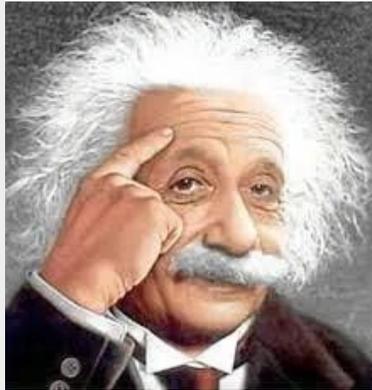
**Edit Request:** "Can Someone remove the building and the white sign from the picture? Please and thank you!"

---

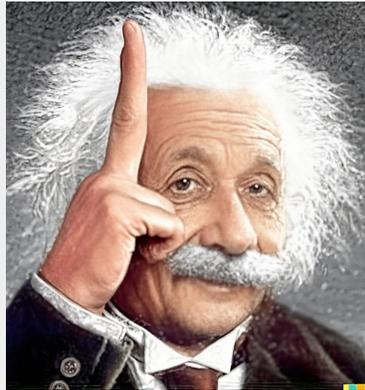
**o1 Verdict:** "A is better"

**o1 Reasoning:** "Both edits successfully remove the large white sign in front, but **only Edit A also removes the conical building** in the background and replaces it with sky. This more completely fulfills the request to remove 'the building and the white sign.' **Edit B leaves the conical structure partially in view**, so Edit A better meets the original instructions."

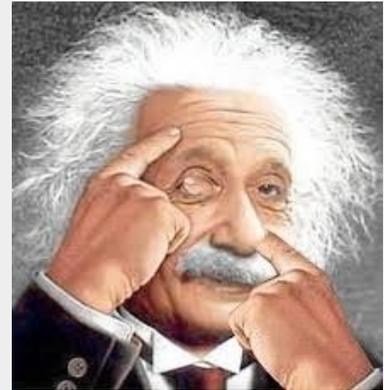
Figure A31. o1 occasionally fails to notice image details when judging different edits. In this example, it thinks only one of the images removed the conical structure, while in reality, both edits removed the building.



Source



Edit A (AI)



Edit B (Human)

**Edit Request:** "Can someone Photoshop this so the finger goes up his nose and out his eye?"

**o1 Verdict:** "A is better"

**o1 Reasoning:** "The user wants a comedic effect of a single finger going in through the nose and emerging from the eye. Edit A extends a finger upward near the nose area, but it doesn't visually pass through to the eye. Edit B mostly shows both hands near the face **without a clear sense of the finger going in the nose and out the eye**. Between the two, Edit A at least attempts to reposition the finger upward near the nose, though it doesn't convincingly show it exiting through the eye. **Edit B doesn't obviously depict the requested effect at all**. Therefore, Edit A more closely matches the user's intent, even if it's not fully realistic."

Figure A32. o1 judge occasionally fails to notice details in edited images, here, overlooking the position of the hand and the configuration of the fingers.

## G.5. Analysis of VLM-as-a-Judge’s Misalignment with Human Preference

Table A21. We conduct a manual analysis of 100 cases where all three VLMs judges disagree with human judges to identify common patterns. The table summarize the main AI judge failure modes.

<b>Failure mode</b>	<b>Count</b>	<b>%</b>	<b>Explanation</b>
Change in subject identity	18	18	Human judges were highly sensitive to changes in the identity or appearance of a person or animal, whereas VLMs often failed to detect these shifts.
Excessive edit	15	15	VLMs tended to favor more dramatic edits, including significant background changes, bright colors, and unnecessary object modifications.
Hallucination	9	9	In several cases, the VLMs inferred or imagined details that were not actually present in the image.
Misunderstood prompt	8	8	Some VLMs failed to correctly interpret the intent behind the editing request.
Watermark bias	2	2	Both edits were technically valid, but the VLMs based their preference on the presence of a watermark in one of the images.
Other / Subjective or judge errors	48	48	The rest of the human-AI judge disagreement included judge errors and genuine subjective differences (where both edits were valid, it is subjective to choose one vs. the other).

## G.6. Qualitative Analysis of AI Wins and Losses to Identify Patterns

**When AI Outperforms Human Preference (206 top-rated edits)** AI wins primarily due to more accurately reflecting user requests:

- Overall, 72% of winning AI edits align closely with user instructions.
- Breakdown by model:
  - GPT-4o: 73% of 62 images
  - Gemini-2.0-Flash: 72% of 32 images
  - SeedEdit: 74% of 76 images
  - Hugging Face models (🤖): 69% of 36 images
- **Common Patterns of AI Failures (400 sampled losses)** Primary reasons for AI losing include:
  - **Misunderstanding user prompts (43% overall):**
    - GPT-4o: 16%
    - Gemini-2.0-Flash: 58%
    - SeedEdit: 46%
    - Hugging Face models (🤖): 50%
  - **Unwanted artifacts or distortions:**
    - Facial identity distortions were prominent in GPT-4o (78%) compared to others:
      - \* Gemini-2.0-Flash: 14%
      - \* SeedEdit: 23%
      - \* Hugging Face models (🤖): 12%
    - Other unrequested changes were found in 14% to 23% of edits across models.

**Overall Insight** AI models excel when accurately interpreting user requests, yet significant challenges remain in understanding instructions and minimizing unintended modifications. The severity of these issues varies significantly across different AI models.

## G.7. AI Edit Observations

We qualitatively observe that many of the models - in particular text-image models like SeedEdit or CosXL - frequently add irrelevant changes to the original image. We manually inspect the top 50 highest win rate AI edited images and find that 22% of them contain edits that are irrelevant to the original edit request.

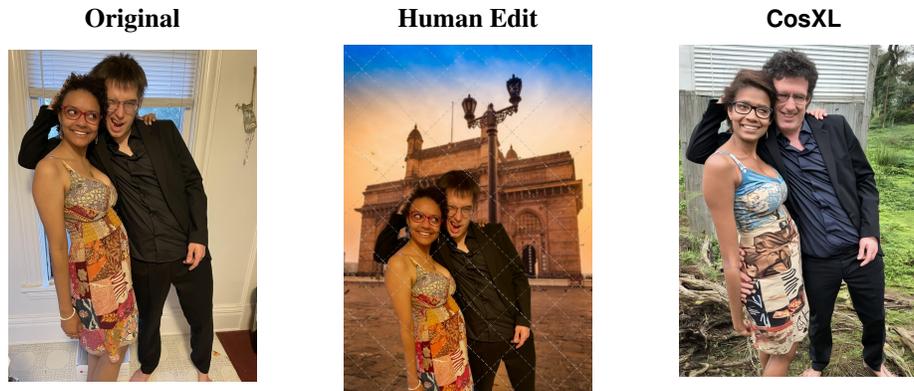


Figure A33. This edit adds unnecessary changes to the peoples' face, hands and clothes despite only needing to edit the background. *User request: Place the two individuals in a different background landscape.*

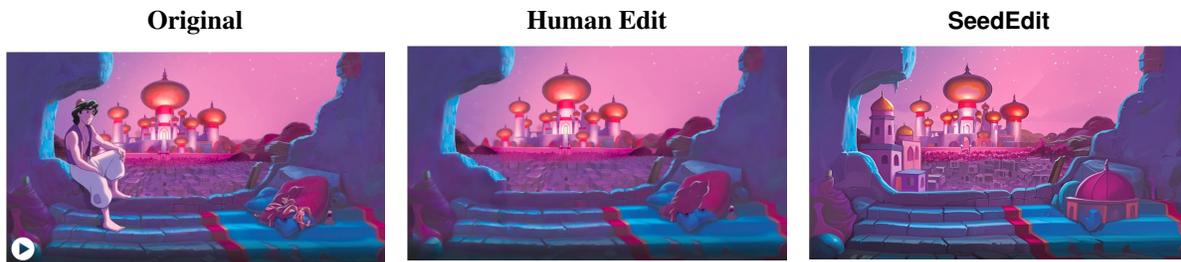


Figure A34. The request only asks for portions of the image to be removed, but the model adds additional buildings to the city background. *User request: Remove Aladdin, Abu, and the play button to focus on the city-scape.*



Figure A35. The model changes the structure of the ring even though the request says to only change the color. *User request: Change the color of the green stones to pale yellow.*

## H. Human Evaluation Sample Size in Related Work

Paper	# of Original Photos	# of Human Participants	Votes
Imagic (CVPR 2023)	100	–	9,213
Re-Inpainting Self-Consistency Eval (arXiv 2024)	100	100	10,000
HIVE (CVPR 2024)	200	3 MT per image	3,000
EditVal Benchmark (ICLR 2024)	92	3 MT per image	15,552
EditBench (CVPR 2023)	240	18	11,520
<b>PSR-328 (Ours)</b>	328	122	4,359

Table A22. Comparison of sample size in related image editing benchmark works.

## I. Additional Samples



Figure A36. Models fail at edits requiring text replacement. User request: *Replace 'SONIC' with 'ROAD RUNNER' and 'HEDGEHOG' with 'BIRD'.*



Figure A37. Models succeed at recoloring and restoring images. User request: *Perform color correction and remove spots from the slide scan.*



Figure A38. Models can recolor specific parts of images. User request: *Change the color of the green stones to pale yellow.*

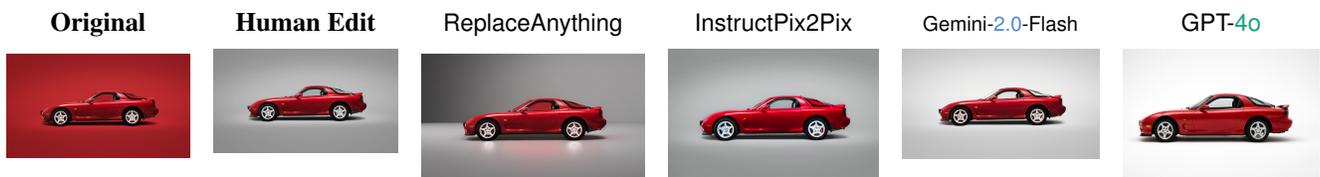


Figure A39. Models can usually change the background properly. User request: *Change the background to white or light grey to simulate a studio setting.*



Figure A40. Models can handle requests to replace objects. User request: *Replace the metal crate at the bottom of the image with additional sidewalk.*



Figure A41. Models can add objects to images, including as part of another object. User request: *Add a pink tutu dress to the elk in the image.*



Figure A42. Models excel at removing the backgrounds of images. User request: *Create a cutout of the raccoon and enhance the image quality to HD.*



Figure A43. Models can enhance images by removing blur. User request: *Remove blur from the subject while keeping the background intact.*



Figure A44. Models succeed at edits that do not require high amounts of realism. *User request: Transform the sky to appear psychedelic.*

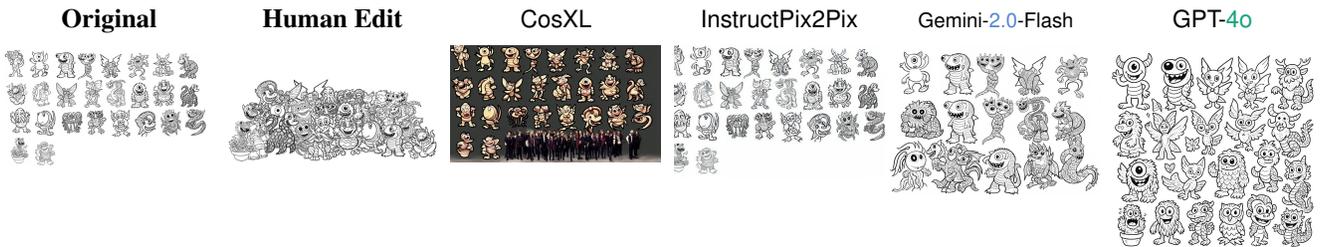


Figure A45. Models fail at edits that require manipulating several objects within the same image. *User request: Arrange the characters into a group picture, resizing as necessary and omitting a few for aesthetics.*



Figure A46. Human users often provide rough reference images as a guide for the type of image they want to create. Models typically struggle to understand these reference images when they are not exact representations of what the final output should appear like. *User request: Edit the android image to appear as if it is drinking from a Java cup.*



Figure A47. Models struggle with requests requiring humor. *User request: Photoshop the soda into humorous scenarios where people's hands are doing something focused or with extreme concentration.*

## J. Generative AI models fail to preserve identity in image editing

We generate an initial image of a person wearing a white T-shirt across various ages and genders using FLUX-Pro [27]. These images are then processed using GPT-4o and Gemini-2.0-Flash to apply different shirt colors. At each step, we send the image from the previous step along with a new color name, repeating this process for eight steps. Finally, we instruct the model to revert the shirt color back to white (the original color). At each step, we compute the  $L_2$  distance between the DINOv2\_ViT-B/14 [41] feature representations of the modified images and the original image (white shirt) to quantify feature deviations.



Figure A48. GPT-4o fails to preserve the identity of individuals in image editing tasks. When tasked with changing the color of a shirt in a sequence, the identity of the person changes—after a few iterations, the person loses their likeness to the original image. The  $d$  value indicates the  $L_2$  distance between the DINOv2\_ViT-B/14 embedding of the image and the original one.



Figure A49. Gemini-2.0-Flash fails to preserve the identity of individuals in image editing tasks. When tasked with changing the color of a shirt in a sequence, the identity of the person changes—after a few iterations, the person loses their likeness to the original image. The  $d$  value indicates the L2 distance between the DINOv2\_ViT-B/14 embedding of the image and the original one.



Figure A50. Although changing the shirt color while preserving every other detail in the image affects the DINO embeddings, the magnitude of the change is much smaller than that caused by any edits made by AI models such as GPT-4o and Gemini-2.0-Flash. In this figure, we change the shirt color using an image editing tool while keeping all other elements unchanged. We report the variation in the DINOv2\_ViT-B/14 embedding compared to the original image.

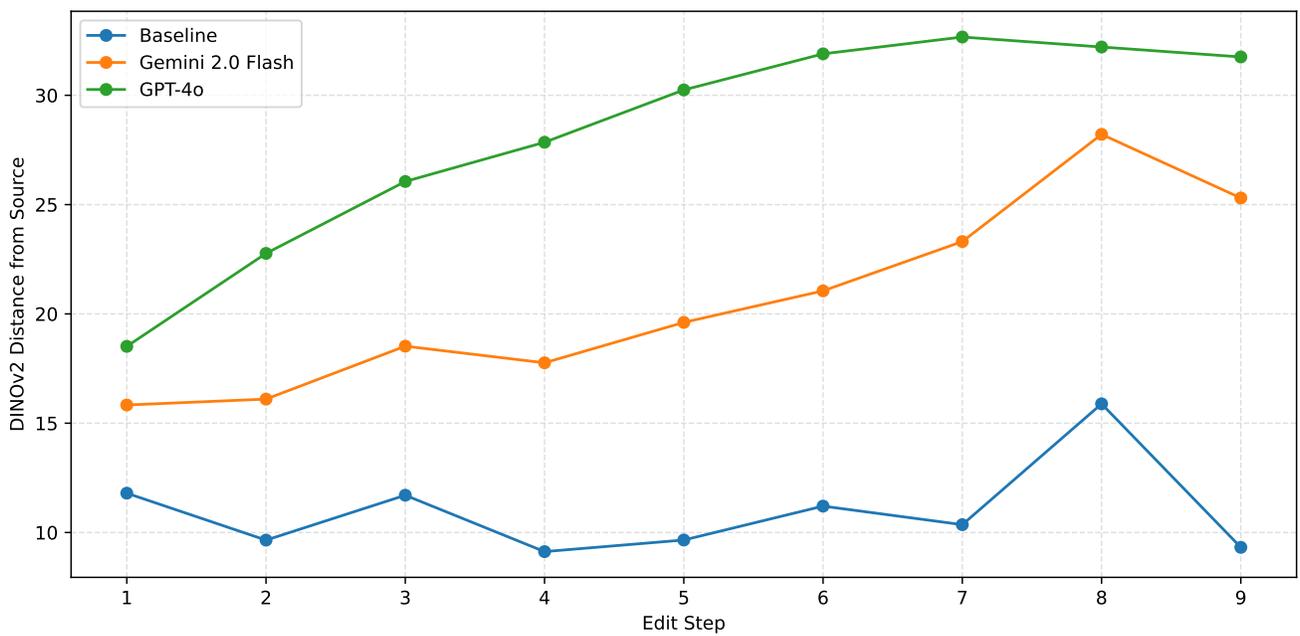


Figure A51. In a sequential image editing request, where a model is subsequently asked to change a shared color, the distance between the edited image at step  $n$  and the original image increases as the requests continue. The value for each data point step indicates the L2 distance between the DINOv2\_ViT-B/14 embedding of the image and that of the original image.

## **K. Dataset License**

All data collected in this study comes from publicly available Reddit posts, which fall under Reddit's user agreement policy.

GPT-4o image edits are produced using the ChatGPT website.

Gemini-2.0-Flash image edits are produced using the standard API.

SeedEdit images edits are generated by the SeedEdit team.

Other AI generated edits were created through the Hugging Face (🤖) website.