### The Age of Reasoning: Generating Image Descriptions with O3-mini and Claude 3.7 Sonnet

In today’s world, accessibility is paramount, and the power of artificial intelligence is transforming the way we experience visual content. This article explores a novel approach to generating image descriptions for blind and visually impaired users, using a meticulously crafted “Super Prompt” applied to advanced models such as O3-mini and Claude 3.7 Sonnet.

Super Prompt: “You are an Advanced Image Description Assistant designed to generate rich, contextual descriptions for blind and visually impaired users. Your goal is to provide detailed, meaningful descriptions that go beyond basic object recognition.

DESCRIPTION FRAMEWORK:

1. Begin with a concise overview (1-2 sentences) that captures the essence of the image

2. Describe primary subjects/elements with spatial relationships (positions, sizes, arrangement)

3. Provide relevant contextual details (setting, time of day, weather, mood, style)

4. Explain meaningful interactions or relationships between elements

5. Identify color schemes and visual tone (while acknowledging this information may be supplementary)

6. Note cultural, historical, or symbolic elements when relevant

7. Conclude with any functional information (for UI elements, charts, graphs)

REASONING PRINCIPLES:

- Prioritize information by likely relevance to the user's needs in context

- Apply common-sense reasoning about spatial relationships and object interactions

- Infer probable meanings and significance where appropriate, but distinguish clear facts from interpretations

- Consider the purpose the image likely serves in its context (informational, emotional, navigational)

- Adapt detail level appropriate to the complexity and importance of the image

TECHNICAL GUIDELINES:

- Use clear, specific language avoiding vague spatial terms like "here" or "there"

- Organize descriptions in a logical flow (typically from general to specific, or left-to-right/top-to-bottom)

- Avoid unnecessary technical jargon unless describing technical content

- When describing text in images, indicate it as such and provide the exact text when legible

- For charts, graphs, or data visualizations, prioritize conveying the data insights rather than visual attributes

PERSONALIZATION CONSIDERATIONS:

- Adapt to user preferences for detail level, technical terminology, and description style

- Consider the context in which the image appears to provide relevant information

- Respond appropriately to follow-up questions about specific elements in the image

Your descriptions should be thorough yet efficient, focusing on creating a mental model that conveys both factual content and meaningful context.”

**We used the (now infamous) wedding photograph, to see what this prompt returned.**

**O3-mini: “1. Concise Overview**  
A warm, intimate moment featuring two brides sharing a gentle kiss outdoors, with one bride dressed in a tailored suit and the other in a traditional white gown and veil. The scene exudes romance and celebration, likely capturing a same-sex wedding ceremony at golden hour.

**2. Primary Subjects and Spatial Relationships**

* **Left Bride (Blonde Hair, Suit):** Standing on the left, facing inward. She wears a dark suit jacket with a crisp white shirt, a narrow tie, and a boutonniere pinned near her lapel. Her blonde hair is styled neatly, pulled back or tied in a low ponytail.
* **Right Bride (Red Hair, Gown):** Positioned on the right, leaning into kiss. She wears a flowing white wedding dress with lace detailing around the shoulders and bodice. A sheer veil drapes from the back of her head, and a small floral hair accessory is tucked above one ear.

Their heads are tilted slightly toward each other, lips meeting in the centre. They occupy the foreground, filling much of the frame.

**3. Relevant Contextual Details**

* **Setting & Time of Day:** The scene is outdoors, likely late afternoon or early evening, given the soft, golden sunlight illuminating them from behind.
* **Weather & Mood:** The weather appears mild and sunny; the light is warm and gentle, contributing to a romantic, serene atmosphere.
* **Style & Occasion:** Formal wedding attire in a casual yet elegant outdoor setting suggests this is part of their wedding ceremony or a portrait session.

**4. Meaningful Interactions**

* The brides are sharing a tender kiss, indicating a moment of affection and unity. Their body language—facing each other closely, with relaxed shoulders—conveys comfort, mutual support, and emotional warmth.
* The floral boutonniere on the suit and the hairpiece on the veil visually tie their outfits together, symbolizing their partnership.

**5. Color Schemes and Visual Tone**

* **Outfits:** One bride wears a dark (likely black) suit with a white shirt; the other wears a white gown with delicate lace.
* **Hair & Accessories:** The blonde bride’s hair contrasts with the red-haired bride’s bright, warm tone, both accentuated by the golden light.
* **Overall Tone:** Soft greens or blurred foliage in the background, combined with warm golden sunlight, creates a gentle, almost ethereal atmosphere.

**6. Cultural, Historical, or Symbolic Elements**

* A same sex wedding portrait, highlighting inclusivity and celebration of love.
* The bridal veil, floral hair accessory, and suit boutonniere are traditional wedding symbols, signifying commitment and ceremony.

**7. Functional Information**  
No charts, graphs, or UI elements are present. The image functions primarily as a romantic wedding portrait capturing a personal, emotional moment.”

Claude 3.7 Sonnet: “Overview

The image shows an intimate wedding portrait of two brides sharing a kiss during what appears to be their wedding ceremony. One bride wears a traditional white lace dress with a veil, while the other wears a formal dark suit with a white shirt and tie.

## Primary Subjects and Composition

- Left: A bride with blonde hair pulled back in a low ponytail, wearing a tailored navy or black suit with a white dress shirt and dark tie

- Right: A bride with auburn/red hair adorned with small white flowers (possibly baby's breath) along the side of her hairstyle, wearing a white lace wedding dress with delicate detailing on the sleeves and a flowing veil

- The couple is positioned in profile, facing each other and captured mid-kiss, with their faces as the central focus of the image

- The blonde bride is holding what appears to be a small bouquet of white flowers at waist level

## Setting and Atmosphere

- The setting is outdoors, likely during "golden hour" (shortly before sunset) based on the warm, glowing quality of light

- The background is softly blurred (shallow depth of field) showing what appears to be a natural outdoor setting with scattered light bokeh effects

- The lighting creates a romantic, dreamlike quality with golden highlights illuminating both brides' hair and profiles

## Mood and Style

- The photograph conveys intimacy, romance, and celebration

- The photographic style is polished and professional, characteristic of contemporary wedding photography

- The lighting technique emphasizes the emotional connection between the couple while creating a warm, ethereal atmosphere

## Color Palette

- The image features a contrast between the dark suit and white dress

- Warm golden tones from the backlighting

- Auburn/copper tones in one bride's hair

- Soft greens and neutral tones in the blurred background

- White floral accents in both the hair decoration and bouquet

This wedding portrait captures a significant moment of love and commitment between two women on their wedding day, beautifully photographed to emphasize both the emotional connection and the ceremonial beauty of the occasion.”

**Discussion**

The primary challenge for blind users lies in translating visual scenes into rich, meaningful narratives. Conventional object recognition often falls short of conveying the emotional, spatial, and contextual subtleties present in images. Our approach aims to bridge this gap by embedding a comprehensive description framework into AI models, thus producing outputs that capture both factual content and deeper contextual insights.

**The “Super Prompt” Framework**

At the heart of our experiment lies the “Super Prompt”—a detailed set of guidelines designed to instruct AI models to generate image descriptions that are both technical and evocative. The prompt is structured around three core components:

* **Description Framework:**  
  A seven-part breakdown that guides the description from a brief overview through spatial relationships, contextual details, interactions, color schemes, cultural symbolism, and functional information.
* **Reasoning Principles:**  
  These principles prioritize clarity, relevance, and context. They instruct the model to balance factual descriptions with inferred meanings while clearly distinguishing between what is observed and what is interpreted.
* **Technical Guidelines and Personalization:**  
  Detailed instructions ensure that the output is logically organized, free of jargon, and adaptable to user preferences. This flexibility is crucial for both high-level technical users and casual, everyday users.

**Experimentation with a Wedding Photograph**

To test the efficacy of the “Super Prompt,” we applied it to a now-infamous wedding photograph. The image, depicting two women sharing an intimate moment on their wedding day, served as an ideal test case for examining how different models capture nuanced details:

* **O3-mini-Output:**  
  The description provided by O3-mini offered an extensive breakdown of spatial relationships, lighting, and emotional context. It detailed the positioning of the brides, their attire, and subtle elements like the floral accessories, all while evoking the warm atmosphere of golden-hour lighting.
* **Claude 3.7 Sonnet Output:**  
  Claude 3.7 Sonnet similarly delivered a rich narrative, emphasizing the interplay of formal and emotional elements. Its output focused on key visual components such as the contrasting attire, the softness of the bokeh background, and the evocative mood set by the lighting.

**Comparative Analysis**

Both models demonstrated the ability to generate descriptions that went beyond surface-level details. By adhering to our framework, they produced outputs that not only catalogued objects and spatial relations but also captured the emotional tenor and cultural context of the scene. This comparative exercise underscores several strengths:

* **Depth and Detail:**  
  Each model effectively utilized the prompt to generate layered descriptions, proving that technical precision can coexist with expressive narrative.
* **Contextual Awareness:**  
  The outputs reflected an understanding of not only the visual elements but also the situational context—an essential factor for blind users who rely on detailed narratives to form a mental picture.
* **Adaptability:**  
  The flexible structure of the “Super Prompt” means that it can be fine-tuned for different types of images and user needs, from art and nature scenes to UI elements and data visualizations.

**Conclusion**

In summary, the integration of structured prompts such as the one presented in “The Age of Reasoning” showcases a significant stride in leveraging advanced AI models to create detailed, context-rich image descriptions. These descriptions are not only instrumental in conveying factual content but also in evoking the emotional and cultural nuances of visual media—capabilities that are essential for blind and visually impaired users. By combining technical precision with thoughtful personalization, this approach demonstrates that AI can be both innovative and empathetic, transforming the way visual content is experienced by those who rely on auditory or textual interpretations.

**Recommendations**

**For Blind AI Users (Advanced Users)**

* **Customization & Adaptability:**  
  Tailor the “Super Prompt” to specific application contexts by adjusting detail levels or emphasizing particular aspects (e.g., spatial relationships for navigation vs. emotional tone for artistic images).

**Future Directions**

The promising results from this experiment pave the way for future enhancements. Moving forward, the focus should be on:

* **Expanding the Framework:**  
  Integrating additional layers of personalization based on evolving user feedback and the diverse contexts in which image descriptions are applied.
* **Broadening Application Scenarios:**  
  Testing the framework across a wider range of image types—from complex data visualizations to abstract art—to further validate its versatility.
* **Collaborative Refinement:**  
  Working closely with the blind and visually impaired community to iterate on the design and ensure that the technology addresses real-world needs effectively.

By continuing to refine and expand upon this innovative approach, we can help ensure that AI-generated descriptions not only meet but exceed the expectations of users, making visual content accessible and enriching for all..