Say Cheese! Detail-Preserving Portrait Collection Generation via Natural Language Edits

Supplementary Material

8. Dataset Statistics and Analysis

To comprehensively evaluate the CHEESE dataset, we conduct a statistical analysis focusing on image quality, content diversity, and instruction complexity. As illustrated in Figure 1, our analysis confirms that CHEESE offers a high-quality, diverse, and creatively challenging benchmark for portrait collection generation.

High-Fidelity and Visual Diversity. High-resolution imagery is a prerequisite for professional photobook generation, ensuring that synthesized details remain sharp and realistic. Figure 8(a) presents the distribution of image heights. Notably, the dataset consists entirely of high-resolution samples, with 40.5% falling within the 1200–1400 pixel range and a substantial 48.9% exceeding 2000 pixels. This ultra-high definition sets CHEESE apart from existing datasets, posing a greater challenge for detail preservation.

To quantify content richness, we employed a LVLM to categorize the images into six predefined scene types and clothing styles. Figure 8(b) shows the scene distribution. While "Studio" settings are predominant (55,189 samples)—reflecting the typical nature of portrait photography—the dataset maintains significant diversity with "Natural Landscape", "Urban Architecture", "Home Setting", and "Commercial Space". Figure 8(c) details the clothing distribution. The dataset covers a wide spectrum of attire, ranging from "Casual" and "Formal/Gown" to culturally specific categories like "Traditional" and "Ethnic" , alongside "Professional" and "Sportswear". This diversity in scenes and garments ensures the model's robustness across varied real-world portrait scenarios.

Instruction Granularity and Creativity. A core contribution of CHEESE is the complexity of its modification instructions, which requires the model to execute creative edits while maintaining reference fidelity. Figure 8(d) illustrates the distribution of modification text lengths, with the majority concentrated between 23 and 33 tokens. This length indicates that the instructions are descriptive and detailed rather than simple, short prompts. We further categorize these instructions into six modification types (Figure 8(e)): Motion Pose, Object , Camera Distance, Viewpoint Angle, Expression, and Light. The dominance of pose and object modifications highlights the dataset's focus on dynamic character interactions.

To analyze the creativity of these requests, we break down each modification type into specific "intents" (Figure 8(f)). For instance, within the Motion Pose category, "Hand Gesture" accounts for 46.6% of the intents, involving intricate actions such as "holding a cup," "writing," "waving," or "pointing." This hierarchical annotation demonstrates the fine-grained control required by the dataset.

Task Complexity and Compositionality. Real-world editing requests often involve changing multiple attributes simultaneously. Figure 8(g) and (h) analyze the compositionality of the modification texts. Figure 8(g) shows that a single text prompt typically triggers 1 to 3 distinct modification types (e.g., changing both the lighting and the pose). More importantly, Figure 8(h) reveals that at a finer granularity, a single prompt usually encompasses 3 to 5 specific modification intents. This high density of intents necessitates that the generative model possesses strong instruction-following capabilities to disentangle and execute multiple complex requirements within a single generation pass.

9. Data Pipeline Prompts

9.1. Image Pair Selection

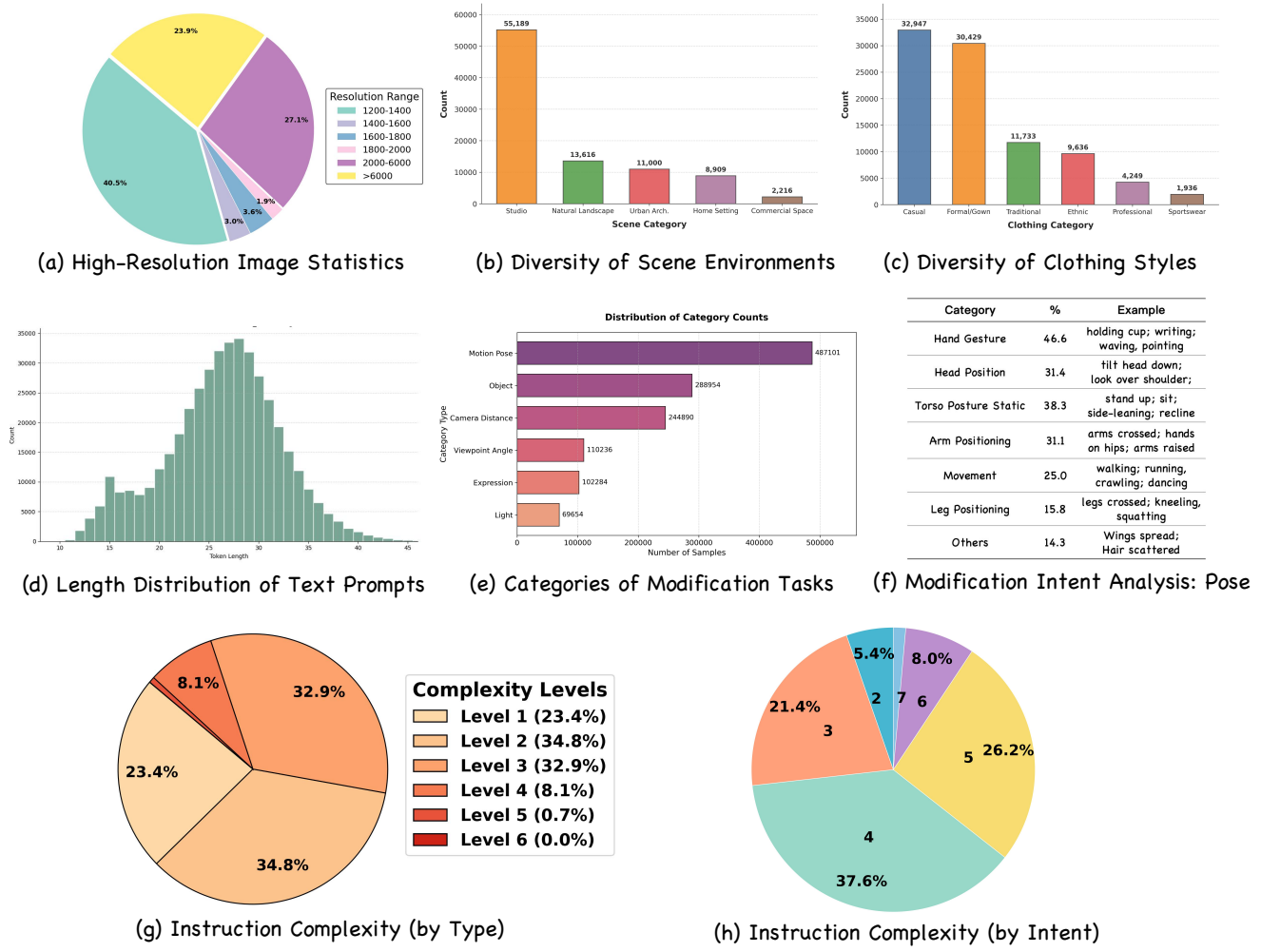
In PCG, effective reference-target image pairs should maintain a certain level of consistency while exhibiting meaningful variations. Specifically, pairs should not introduce substantial unknown or uncontrollable new elements, yet should contain sufficient changes to support meaningful editing. To achieve this balance, we employ an LVLM to filter out near-duplicate pairs and pairs with excessive background changes. The filtering process uses the prompt shown in Figure 9.

9.2. Modification Text Generation

To generate high-quality modification texts, we design a structured prompt, as shown in Figure 10, which guides the LVLM to describe transformations from reference to target images. The prompt emphasizes three key modification aspects: (1) camera and viewpoint changes (framing, distance adjustments); (2) model pose and body orientation (posing, hand positions, expressions, frame position); and (3) object and background changes. The prompt enforces quality constraints including token limits (under 77 tokens), detailed object descriptions, avoidance of vague terms, and use of specific, professional language. During annotation, the LVLM receives iterative feedback from previous attempts to refine the generated instructions.

9.3. Inversion-based Validation

To ensure annotation quality, we introduce an inversion-based verification mechanism. As shown in Figure 11 ,



(a) High-Resolution Image Statistics (b) Diversity of Scene Environments (c) Diversity of Clothing Styles

(d) Length Distribution of Text Prompts (e) Categories of Modification Tasks (f) Modification Intent Analysis: Pose

Category	%	Example
Hand Gesture	46.6	holding cup; writing; waving, pointing
Head Position	31.4	tilt head down; look over shoulder;
Torso Posture Static	38.3	stand up; sit; side-leaning; recline
Arm Positioning	31.1	arms crossed; hands on hips; arms raised
Movement	25.0	walking; running, crawling; dancing
Leg Positioning	15.8	legs crossed; kneeling, squatting
Others	14.3	Wings spread; Hair scattered

(g) Instruction Complexity (by Type) (h) Instruction Complexity (by Intent)

Figure 8. **Dataset Statistics and Analysis of CHEESE.**

Prompt for Image Pair Selection

```
Evaluate whether this portrait image pair should be included in a dataset
for portrait collection generation.

FILTER OUT if:
1. Near-duplicate: Images are almost identical with only minor differences
2. Excessive background change: Background/scene has changed
dramatically (different locations, completely different settings)

KEEP if:
- Meaningful changes in pose, expression, camera angle, or spatial layout
- Moderate or minimal background changes
- Subject identity remains consistent

Respond with "FILTER" or "KEEP".

Here is the images:
```

Figure 9. **Prompt for Image Pair Selection.**

given modification text and reference image, an LVLM generates a predicted target caption describing the expected edited image. The prompt requires: (1) include person details; (2) describe new states for mentioned elements, original states for unmentioned elements; (3) focus on concrete

Prompt for Modification Text Generation – P_m

```
You will see two images: a reference image (first) and a target image (second).
Your task is to describe the changes needed to transform FROM the reference image TO the target image in one
sentence. This is crucial - always describe the transformation starting from the reference image.

1. Focus on changes in:
   - Camera view (ONLY when significant changes are needed):
      * Framing and distance:
        - "Frame for a tight close-up of the face"  - "Pull back for a full-body composition"  - "close-up/half-
body/full-body shot"

   - Model poseture:
      * Posing and Body's direction to the camera
        - "Adjust body to a seated position"  - "Turn to the right"  - "Turn counterclockwise to show left profile"
      * Hand and arm positions:
        - "Place hands folded in lap"  - "Raise right hand to chin level"  - "Cross arms at chest height"
      * Expression and details:
        - "Tilt the model's chin down"  - "Look directly into the lens"  - "Turn head to the left"
      * Position in the frame:
        - "move the model in the middle of the frame"

   - notable objects (e.g., moving an umbrella from above head to behind)
   - new objects (e.g., "fish-shaped lantern")
   - Change of background

2. Important principles:
   - MUST be under 77 tokens
   - Ignore slight changes: When changes are minimal (slightly different), consider them unchanged
   - For new key objects: Provide detailed descriptions (e.g., "fish-shaped lantern" instead of just "lantern")
   - Avoid vague or subjective terms:  Instead of "more/better", provide exact changes needed
   - Use professional but clear language with specific directions

3. Do not mention "reference image" or "target image" in your description.
4. Ensure the description is concise, limited to one sentence.

I will provide your previous answer and their scores (if any). Higher scores indicate better instructions that more
accurately capture the desired changes, and your goal is to generate an instruction rated at 1.0.
Please refer to the following answer and their scores to enrich your current answer:
```

Figure 10. **Prompt for Modification Text Generation.**

visual elements; We then compute CLIP similarity between target image and predicted caption, and regenerate the modification text using the failed attempt as feedback, if the current one is not good enough.

Prompt for Inversion-based Validation – P_I

You are a professional image caption generator. I will provide an image and an editing instruction. Your task is to generate a complete description of how the edited image would look, which will be used to retrieve the edited image from a search engine.

REQUIREMENTS:
1. Keep it under 77 tokens
2. Include the details of the person in the image.
2. Important guidelines:
  - For elements mentioned in the editing instruction, describe their new state
  - For elements not mentioned in the editing instruction, include them as they appear in the original image
  - Focus on concrete visual elements, not abstract concepts
  - Avoid using words like "remain", "add", "lack" in your description
  - Avoid talking about the removed elements

Here is the editing instruction and image:

Figure 11. **Prompt for Inversion-based Validation.**

10. LVLM-based Evaluation Prompts

10.1. Detail Preservation

We employ an LVLM-based metric to assess detail preservation between generated and reference images. As shown in Figure 12, the evaluator determines if images could belong to the same portrait collection, with score 0 assigned to direct copy-paste cases. The evaluation criteria is four aspects: (1) **Model Details:** facial features, skin texture, makeup, hair style; (2) **Outfit Details:** clothing, fabric texture, accessories; (3) **Photography Style:** lighting, color grading, background; (4) **Technical Quality:** sharpness, exposure, natural proportions.

10.2. Prompt Following

As shown in Figure 13, we assess whether generated images accurately implement modification instructions. Evaluation considers: (1) **Implementation Accuracy:** correct application of camera, pose, and accessory changes; (2) **Completeness:** all requested modifications present; (3) **Precision:** matches exact specifications; (4) **Consistency:** no unrequested changes. Scores 0-4: 0 (none correct), 1 (mostly incorrect), 2 (major deviations), 3 (minor deviations), 4 (perfect implementation).

11. Diverse Editing Capabilities of SCheese

We demonstrate SCheese's capability to handle several key editing challenges in Figure 14. **Camera distance variation** includes pulling back for "*half-body shots*" or framing "*tight close-ups*", requiring scale adaptation while preserving facial details and accessories. **Pose transformation** encompasses complex body movements, such as *raising hand* or adjusting to *seated positions with specific hand placements*. **Viewpoint transformation** involves rotating the subject to show different profiles (e.g., "*left/right*"), requiring consistent identity preservation across viewing angles. **Light change.** includes *"starburst effects"* or *"more shadow"*. **Object change.** includes *"adding fabric"* or *"pink bow mirror"*. As shown in Figure 14, SCheese successfully handles all these types of modifications while preserving fine-grained details from the reference image, demonstrating its effectiveness in addressing the diverse editing requirements of PCG.

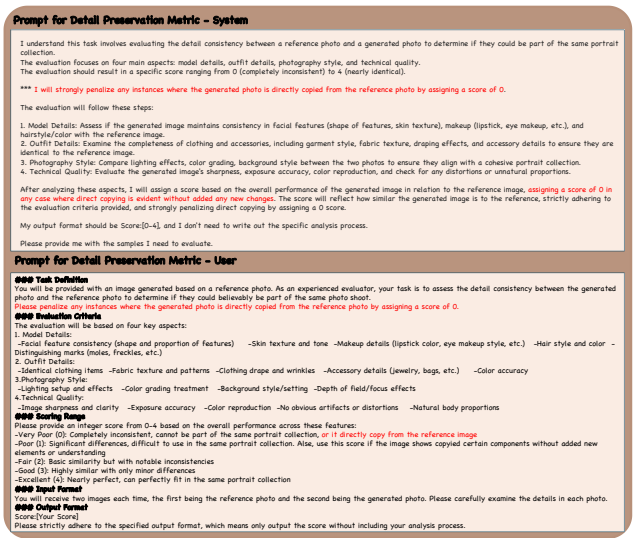


Figure 12. **Prompt for Detail Preservation Metric.**

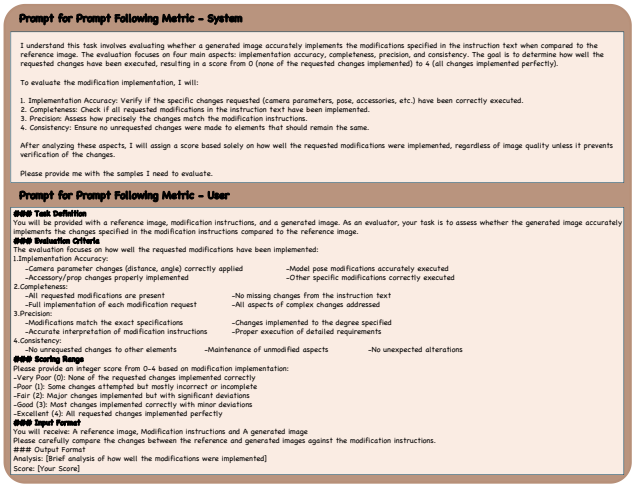


Figure 13. **Prompt for Prompt Following Metric.**

12. Additional Qualitative Results

More qualitative results are provided in Figure 15. As illustrated in the figure, SCheese enables users to generate complete portrait collections with consistent identity and rich

content from a single reference portrait image and multiple modification texts, demonstrating the practical utility of our approach.

13. Data Privacy and Ethical Considerations

To ensure ethical use and protect privacy, we implement comprehensive data anonymization procedures. First, during portrait collection selection, we exclude albums containing minors, pregnant individuals, and religious elements.To protect user privacy, we perform face replacement with synthetic virtual faces on all images.

Dataset Usage Restrictions. The CHEESE dataset is intended solely for research purposes. Users must agree not to use the dataset for commercial purposes, identity recognition, or any activities that could harm individuals. The dataset is provided under strict usage terms that prohibit attempts to reverse-engineer or recover original identities.

Data Security. During dataset construction, all original images are processed in secure environments and permanently deleted after anonymization. Only anonymized versions are retained in the final dataset.

Figure 14. **Qualitative Examples of Diverse Editing Challenges.** SCheese is capable of camera distance, pose variation, and viewpoint transformation, light change and object change while preserving fine-grained details.

Figure 15. **More Qualitative Results.** SCheese can generate complete portrait collections with consistent identity and rich content.