Visual-Aware CoT: Achieving High-Fidelity Visual Consistency in Unified Models

Supplementary Material

The supplementary material is organized as follows:

- **Section 1** introduces implementation details, including the construction of planning and correction datasets, system prompts, and training configurations.
- **Section 2** provides comprehensive experimental results, including more complex multi-reference generation experiments with style reference, and additional ablation studies on reward model design.
- **Section 3** demonstrates qualitative results, showing more iterative refinement examples, detailed comparisons with baseline methods, reasoning process comparisons with text-based CoT methods, and analysis of failure cases.

1. Implementation Details

1.1. Dataset Construction

We construct our planning and correction datasets using a systematic approach based on multi-reference image editing scenarios. Our dataset construction pipeline consists of two main components: planning dataset generation and correction dataset generation.

1.1.1. Data Source and Sampling

We randomly sample 4,000 multi-reference data samples from the Echo-4o dataset [2]. Each sample contains a triplet of reference images, editing instruction, ground truth image, where reference images provide visual context, editing instruction describes the desired editing operation in natural language, and ground truth image represents the expected output after applying the editing instruction.

1.1.2. Planning Dataset Generation

The planning dataset aims to teach the model to decompose complex editing instructions into structured, actionable visual plans. For each sample triplet {reference images, instruction, ground truth image}, we generate an adaptive visual plan that serves as a step-by-step checklist for the editing process.

We employ Gemini-2.5-Flash [1] as our plan generation model, using the system prompt detailed in Figure 1. The generated visual plans contain key visual elements to identify from reference images, specific editing operations to perform, and quality criteria for evaluating the editing results.

1.1.3. Correction Dataset Generation

The correction dataset focuses on training the model's self-evaluation and iterative refinement capabilities. For each sample, we construct a 5-tuple {reference images, instruction, ground truth image, visual plan, current generated image} where the current generated image is generated from a sub-optimal model with the {reference images, editing instruction, ground truth image}.

Using this 5-tuple as input, we employ the system prompt shown in Figure 2 to generate self-evaluation results that provide detailed assessment of the current image quality against the visual plan, and editing instructions that offer specific guidance for improving the current image to better match the target.

```
You generate consistency check instructions for image evaluation. Identify main instances and
format checks according to these rules:

Three Check Types:
1. Identity Consistency:
•Format: "check the identity consistency between the [object/person] in image_X and in
generated image"
2. Attribute Consistency:
•Format: "check the attribute consistency between the [object's attribute] in image_X and in
generated image"
3. Style Consistency:
•Format: "check the style consistency between the artistic style in image_X and in generated
image"

Rules:
•Focus on main instances only (prominent, significant elements)
•Avoid overly specific details (e.g., "left wrist", "small button")
•Choose most relevant consistency type for the context

Examples:
Good:
•"check the identity consistency between the person in image_1 and in generated image"
•"check the attribute consistency between the car's color in image_2 and in generated image"
•"check the style consistency between the artistic style in image_3 and in generated image"
Bad:
•"check the consistency between the person's left wrist in image_1 and in generated image" (too
specific)

Directly generate appropriate checks based on given images/descriptions.
```

Figure 1. **System prompt for visual plan generation.** The prompt guides the model to create structured, actionable plans for image editing tasks.

1.2. Training Sequence

The Bagel team's implementation enables flexible formatting of interleaved image-text data for training through the `add_text` and `add_image` functions. A critical parameter is `need_loss`, which indicates whether ground truth is expected for loss calculation during training.

We demonstrate our training sequence design in Figure 3, where gray blocks represent segments without loss calculation (`need_loss=False`) and blue blocks represent segments requiring loss calculation (`need_loss=True`). Our training data encompasses three distinct types of samples: planning data samples, correction data samples for sub-optimal results, and correction data

```
SYSTEM_PROMPT = """
You are an AI image quality auditor. Your task is to evaluate a generated target image
against reference images and a text instruction and check list.

**Evaluation Process:**

1. Evaluate with the check list and identify the inconsistencies
2. Evaluate overall text-image consistency and identify any unreasonable/impossible
elements
3. Provide editing suggestions towards the ground truth generated image

**Output Format:**
Return a JSON object with this structure:

```json
{
"status": "INCONSISTENT" | "ALL_IS_WELL",
"inconsistencies": {
"text_issues": "<describe text-image mismatches or 'None'>",
"instance_issues": [
{
"instance": "<instance name>",
"reference": "<image_1/image_2/etc>",
"problem": "<describe differences>"
}
]
},
"editing_instructions": "<concise edit commands or 'None if ALL_IS_WELL'>"
}
```

Figure 2. **System prompt for evaluation and correction generation.** The prompt enables the model to assess current results and provide specific editing instructions for improvement.

samples for perfect results.

In the planning data samples, the model learns to generate structured visual plans given reference images and instructions. For correction data samples with sub-optimal results, the model practices identifying discrepancies and providing specific improvement guidance. For correction data samples with perfect results, the model learns to recognize when no further editing is needed and provides appropriate feedback.

1.3. Object Similarity Score

As shown in Figure 4, we compute object similarity scores to measure identity consistency between reference and generated images. We first parse the visual plan checklist to identify target objects. For example, the checklist specifies: "Check the identity consistency between horse in image 1 and in generated image" and "Check the identity consistency between woman in image 2 and in generated image."

We use GroundingDINO to locate these objects in both reference and generated images based on the text descriptions. After extracting bounding boxes for the specified objects, we compute DINO feature similarity between the cropped regions. Higher scores indicate better object identity consistency. This object similarity score provides fine-grained feedback for GRPO training on whether the generated image preserves the visual identity of key objects from the reference images.

To validate whether the object similarity score is reasonable, we test it on our training dataset with both sub-optimal generation results and ground truth images. We find that in 78.3% of the data samples, the ground truth has a higher score than the sub-optimal generation, which means our object similarity score effectively distinguishes between high-quality and low-quality results in most cases.

1.4. Training Configuration

We train VACoT on 8 NVIDIA A800 GPUs using the Adam optimizer with a constant learning rate of 2e-5. We apply a linear warmup schedule that gradually increases the learning rate from zero over the first 500 steps. During training, we randomly sample from our planning and correction datasets, packing sequences to a maximum length of 32,000 tokens per batch. All model parameters except ViT and VAE are jointly optimized throughout the training process.

2. Experiments

2.1. Complete Comparison with other CoT methods

The evaluation of the OmniContext benchmark consists of two dimensions: Prompt Following (PF score) and Subject Consistency (SC score). Due to space constraints in the main paper, we present the detailed scores here. As shown in Table 1, our method consistently outperforms other CoT methods on subject consistency, demonstrating that our design and training approach effectively enhances visual-aware generation capabilities.

2.2. Ablation Study: Reward Model Design

Reward Function Analysis Table 2 presents an ablation study on different reward function configurations for GRPO training. We compare using object similarity score (ObjSimScore) alone, ObjSimScore combined with CLIP Score, and the full combination of ObjSimScore, PickScore, and CLIPScore. The results show that adding CLIPScore to ObjSimScore improves performance across all categories, achieving the best average score of 8.44. However, incorporating PickScore degrades performance, suggesting it may introduce conflicting optimization signals. The optimal reward function combines ObjSimScore and CLIPScore, balancing object-specific similarity with semantic alignment.

Inference Cost versus Benefits. To clarify whether the performance gains come from extra compute, we compare against BAGEL using 3 iterative self-refinement steps. Results in table below show that simply increasing inference cost cannot match our benefits brought by careful designs. Since corrections typically succeed within 3 iterations, the refinement steps in these rounds are highly effective, making the performance gain well worth the inference cost.
```

Figure 3. **Training sequence design for different data types.** Gray blocks indicate no loss calculation, while blue blocks require loss calculation.

Table 1. **Comparison of different multi-modal Chain-of-Thought on OmniContext.**

	Multi Character		Multi Object		Multi Char Obj		Scene Character		Scene Object		Scene Char Obj		Overall
Model	**PF**	**SC**	**PF**	**SC**	**PF**	**SC**	**PF**	**SC**	**PF**	**SC**	**PF**	**SC**	**Average**
UiG	6.90	4.92	8.24	8.72	8.30	8.30	5.54	5.20	6.94	7.60	7.08	7.38	6.986
UniCoT	8.34	6.48	8.92	9.06	8.34	7.96	7.74	6.82	8.12	8.38	8.34	8.10	7.893
Ours	8.98	7.10	9.26	9.26	8.60	8.18	8.27	7.24	8.80	8.60	8.10	8.16	8.257

Method	Iters (N)	Latency	Score ↑	Method	Iters (N)	Latency	Score ↑
BAGEL	1	1.0x	5.55	BAGEL	3	3.1x	6.42
UniCoT	3	3.1x	7.89	UniCoT	5	5.2x	7.54
UiG	3	3.1x	6.85	UiG	5	5.2x	6.37
VACoT	**3**	**3.1x**	**8.44**	**VACoT**	**5**	**5.2x**	**7.70**

Table 2. **Ablation Study on GRPO Reward Functions**. The table shows performance scores across different reward function configurations.

Reward Function	**Character**	**Object**	**Char. + Obj.**	**Average↑**
ObjSimScore	7.45	8.92	7.88	8.08
ObjSimScore + CLIPScore	7.82	9.21	8.30	8.44
ObjSimScore + PickScore + CLIPScore	6.89	8.15	7.42	7.49

2.3. Complex Multi-reference Generation

We evaluate the zero-shot capability of our method on complex multi-reference generation tasks that require both identity preservation and style transfer. As illustrated in Figure 5, the baseline BAGEL model fails to generate images that accurately reflect the target style while maintaining the subject's identity. In contrast, our approach demonstrates robust performance on this challenging task despite not being trained on style transfer data. This zero-shot capability highlights the effectiveness of our method in handling complex multi-reference scenarios where multiple visual attributes must be simultaneously controlled.

3. More Qualitative Results

3.1. Iterative Refinement Examples

We demonstrate the effectiveness of our iterative refinement mechanism through several representative examples. As shown in Figure 6, initial generated images often contain various types of defects, including identity distortion, miss-

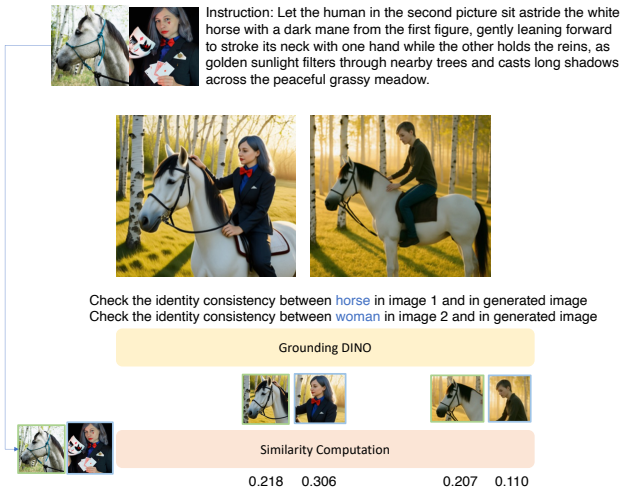


Figure 4. **The object similarity score used in GRPO training.**

ing identity features, or unreasonable elements. Our method systematically identifies these issues and generates appropriate editing instructions to progressively correct them. Through this iterative process, the model can automatically refine the generated images to achieve higher quality results that better preserve the target identity while maintaining visual coherence.

3.2. Comprehensive Method Comparisons

We present extensive qualitative comparisons with other Chain-of-Thought (CoT) based methods to demonstrate the superiority of our approach. As illustrated in Figure 7, our method consistently achieves higher subject consistency and superior identity preservation compared to existing CoT-based approaches. The visual results clearly show that our method maintains better facial features, expressions, and overall identity characteristics while generating contextually appropriate images.

3.3. Ablation Study on Training Stages

We conduct a qualitative ablation study to analyze the contribution of different training stages in our method. Specifically, we evaluate three configurations: our full method, a variant without Supervised Fine-Tuning (SFT), and a variant without GRPO. The results are presented in Figure 8.

The results reveal that both training stages contribute to the overall performance, but with different levels of importance. SFT plays a more critical role in establishing the foundational capabilities of the model, as evidenced by the significant performance degradation when it is removed. GRPO provides additional refinement and optimization. The combination of both training stages achieves the best performance, demonstrating that our two-stage training strategy effectively enhances the visual-aware ability of the model.

3.4. Reasoning Process Comparison

We conduct a detailed comparison of our reasoning process against UniCoT to highlight the advantages of our approach. Starting with a sub-optimal initial result generated by UniCoT, we demonstrate the self-evaluation capabilities of both methods in Figure 9. UniCoT's reasoning process focuses on text alignment with the user prompt, checking only for the presence of specified objects (e.g., whether a pepper exists in the image) while overlooking critical visual attributes such as identity-specific features. In the shown example, UniCoT correctly identifies the presence of a pepper but fails to recognize that the pepper's color does not match the specified requirements, thus missing a fundamental aspect of the prompt. In contrast, our method employs visual-aware reasoning that can accurately identify such discrepancies and provide appropriate corrections.

3.5. Failure Case Analysis

We find that problems that can be fixed are usually resolved within 2 iterations. Beyond this point, the remaining issues become much harder to correct, and additional iterations often make things worse.

Our analysis shows two main reasons why more iterations fail. First, repeated editing degrades image quality by accumulating noise and artifacts with each modification. Second, the problems that persist after 2 iterations are typically fundamental issues that our method struggles to identify correctly or address effectively. When the model tries to fix these harder problems, it often misdiagnoses the issue or applies inappropriate corrections.

Figure 10 demonstrates these failure cases. During the 5 iterations shown, each iteration identifies a new problem, but these problems are not important or meaningful. Instead of improving the result, attempting to fix these minor issues makes the final output worse.

References

[1] Google. Gemini 2.0 flash, 2025. 1

[2] Junyan Ye, Dongzhi Jiang, Zihao Wang, Leqi Zhu, Zhenghao Hu, Zilong Huang, Jun He, Zhiyuan Yan, Jinghua Yu, Hongsheng Li, et al. Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation. *arXiv preprint arXiv:2508.09987*, 2025. 1

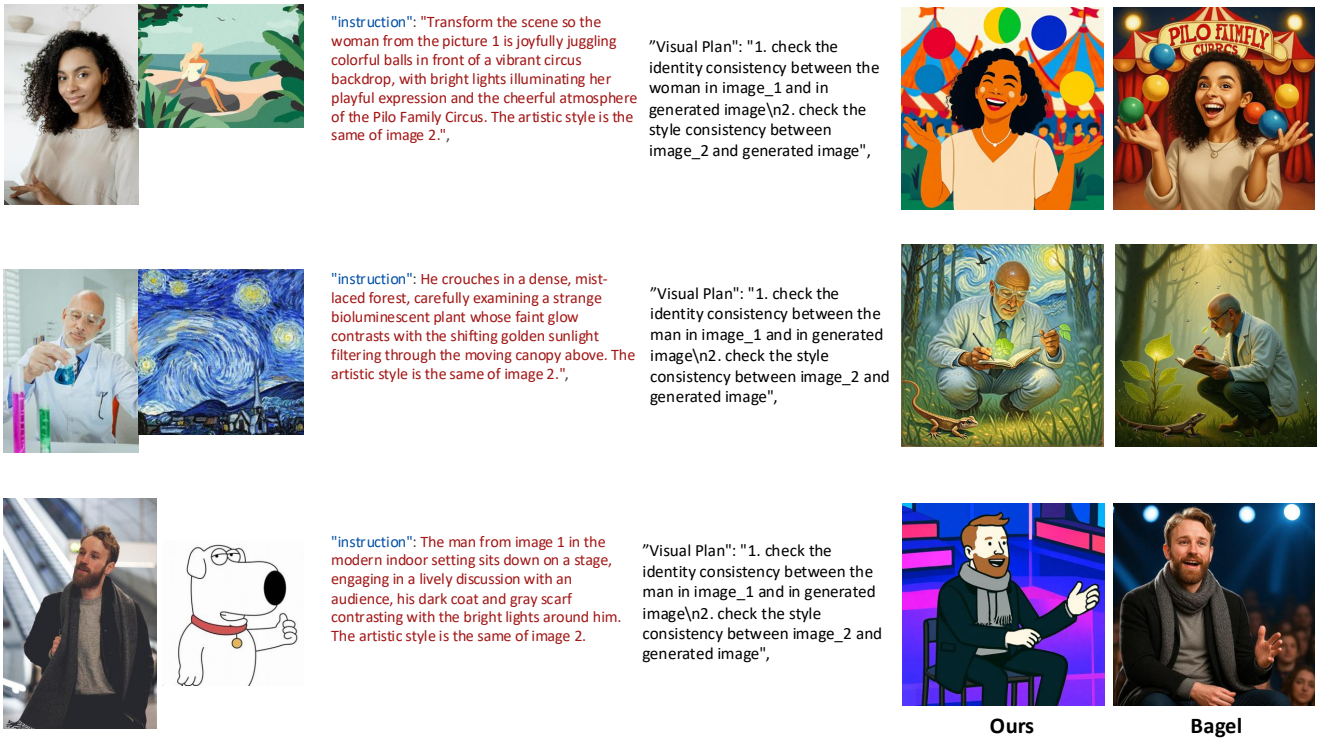
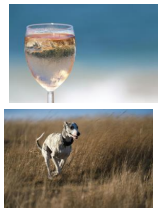


Figure 5. **Results on complex multi-reference generation.**



A wine glass filled with a light-colored beverage rests on a rustic wooden dining table, while a Greyhound stands nearby, its curious gaze directed towards the warm and inviting atmosphere of the room.

"1. check the identity consistency between the wine glass in image_1 and in generated image\n2. check the identity consistency between the Greyhound in image_2 and in generated image"

Place the wine glass on a rustic wooden table. Adjust the dog's position to be more natural. Change the background to show a warm, inviting room atmosphere instead of a field."

The bouquet in the image is positioned on the marble pedestal beneath the grand columns.

"1. check the identity consistency between the flower in image_1 and in generated image

Replace the flower with the flowers from image_1.

Please make the man touch the flower bud, his fingers lightly resting on the textured, white and purple striped bud, with water droplets glistening on its surface.

"1. check the identity consistency between the man in image_1 and in generated image\n2. check the identity consistency between the flower in image_2 and in generated image"

Add the man from image_1 with his finger lightly resting on the purple bud.

The man carries the wolf on his shoulders through a dense forest path, his muscles tense as he strides steadily over uneven ground. Dappled sunlight filters through the thick canopy above, casting shifting patterns of light and shadow on their forms. The wolf rests calmly, its fur blending with the rugged textures of bark and leaves around them.

"1. check the identity consistency between the man in image_1 and in generated image\n2. check the identity consistency between the wolf in image_2 and in generated image",

Change the man's posture to carry the wolf, make the man's face visible and the same as image_1.

Figure 6. **Iterative refinement during generation.**

Figure 7. **Qualitative comparison with other CoT-based methods on OmniContext.** Our method demonstrates superior identity preservation and subject consistency across diverse generation scenarios compared to existing approaches.

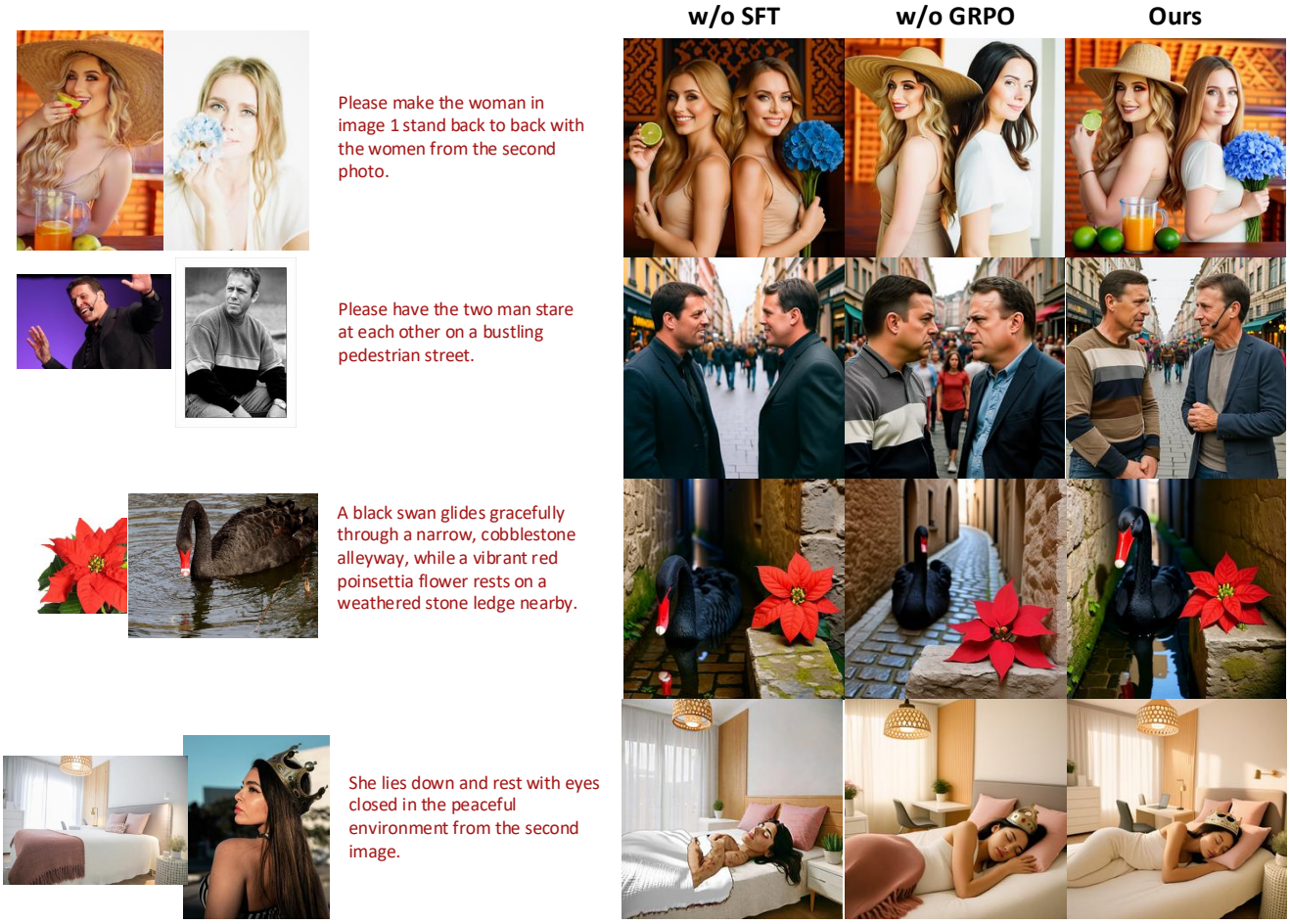


Figure 8. **Ablation study on training stages.** Comparison of our full method against variants without SFT and without GRPO, showing the relative importance of each training component.

Place the pink rose and the pepper on the bedside table in the cozy cabin interior.

Sub-optimal results from UniCoT as the start of CoT

UniCoT

<eval>
1) **Key Expectations and Real-World Knowledge:** - The image should depict a cozy cabin interior. - A bedside table should be present. - A pink rose and a pepper should be placed on the bedside table. - The setting should be consistent with a cozy, rustic environment. 2) **Checklist:** - Is the setting clearly a cozy cabin interior? - Is there a bedside table in the image? - Are a pink rose and a pepper present on the bedside table? - Does the overall composition maintain a rustic and cozy atmosphere? 3) **Assessment:** - The setting is a cozy cabin interior with wooden walls and a snowy forest view, which aligns with the cozy, rustic atmosphere. - There is a bedside table near a window. - A pink rose is present on the bedside table, but a pepper is missing. - The absence of a pepper does not significantly impact the overall cozy atmosphere.The image successfully depicts a cozy cabin interior with a bedside table and a pink rose. However, the pepper is missing from the bedside table as specified in the prompt.
</eval>
<sugg>
Add a pepper to the bedside table alongside the pink rose to fully align with the prompt.
</sugg>

<eval>
1) **Interpretation of the Prompt:** - The prompt requires the placement of a pink rose and a pepper on a bedside table within a cozy cabin interior. - The setting should reflect a cozy atmosphere, typical of a cabin, which might include wooden elements and a warm ambiance. - The bedside table should be identifiable and appropriately positioned within the scene. 2) **Checklist for Evaluation:** - Is there a pink rose present on the bedside table? - Is there a pepper present on the bedside table? - Is the bedside table correctly identified and positioned in the scene? - Does the scene reflect a cozy cabin interior with appropriate elements (e.g., wooden furniture, warm lighting)? - Is the snowy forest view visible through the window, maintaining the cabin's setting? 3) **Assessment:** - The pink rose is present on the bedside table. - The pepper is present on the bedside table. - The bedside table is correctly positioned in the scene. - The scene effectively reflects a cozy cabin interior with wooden elements. - The snowy forest view is visible through the window, aligning with the cabin setting.The image accurately fulfills the prompt by displaying a pink rose and a pepper on a bedside table within a cozy cabin interior, with the correct spatial arrangement and ambiance.
</eval><sugg>
Everything is good. No editing needed.
</sugg>

VACoT

The image does not meet the requirements. The generated image shows a pink rose on a bedside table in a cozy cabin interior, but there is no orange pepper in image_3 present. The original prompt specified the placement of both the pink rose and the orange pepper on the bedside table. Therefore, the image is missing the orange pepper, which is a discrepancy. To correct this, the orange pepper should be added to the bedside table in the generated image.
EDIT_NEEDED: Add an orange pepper from image_3 to the bedside table in the generated image.

The image matches the prompt well. The pink rose and orange pepper are correctly placed on the bedside table in a cozy cabin interior with a snowy forest view. The visual details, color, and environment are consistent with the prompt. The overall mood and atmosphere are also matching the intended description. All identity consistency check items pass verification.
ALL_IS_WELL

Figure 9. **Reasoning comparison with UniCoT.**

Please make the girl whisper to the women at a quiet intersection.

Fix the characters to show whispering, change setting to quiet intersection, replace with correct Healthometer professional scale, and adjust lighting to match outdoor environment.

1. Replace the background with an urban intersection setting. 2. Change the girl's attire to something more suitable for an intersection scene. 3. Add elements to create a quiet atmosphere, such as a calm street and minimal background activity. 4. Adjust the poses of the individuals to suggest a whispering interaction."

remove the healthometer scale from the image.

Change the setting to a quiet intersection. Add the girl from image 1 whispering to the women from image 2. Ensure the women are present and reacting appropriately. Adjust the environment to depict an intersection with appropriate street elements. Ensure the clothing and style remain consistent with the original images."

1. Change the girl's uniform to the one from the first image. 2. Add a headscarf to the woman to match the style in the first image. 3. Change the setting to a quiet intersection, not a busy one. 4. Adjust the girl's hair to be long and braided. 5. Modify the woman's expression to be more serene and intimate."

Figure 10. **Failure case analysis.**