HiCoGen: Hierarchical Compositional Text-to-Image Generation in Diffusion Models via Reinforcement Learning

Supplementary Material

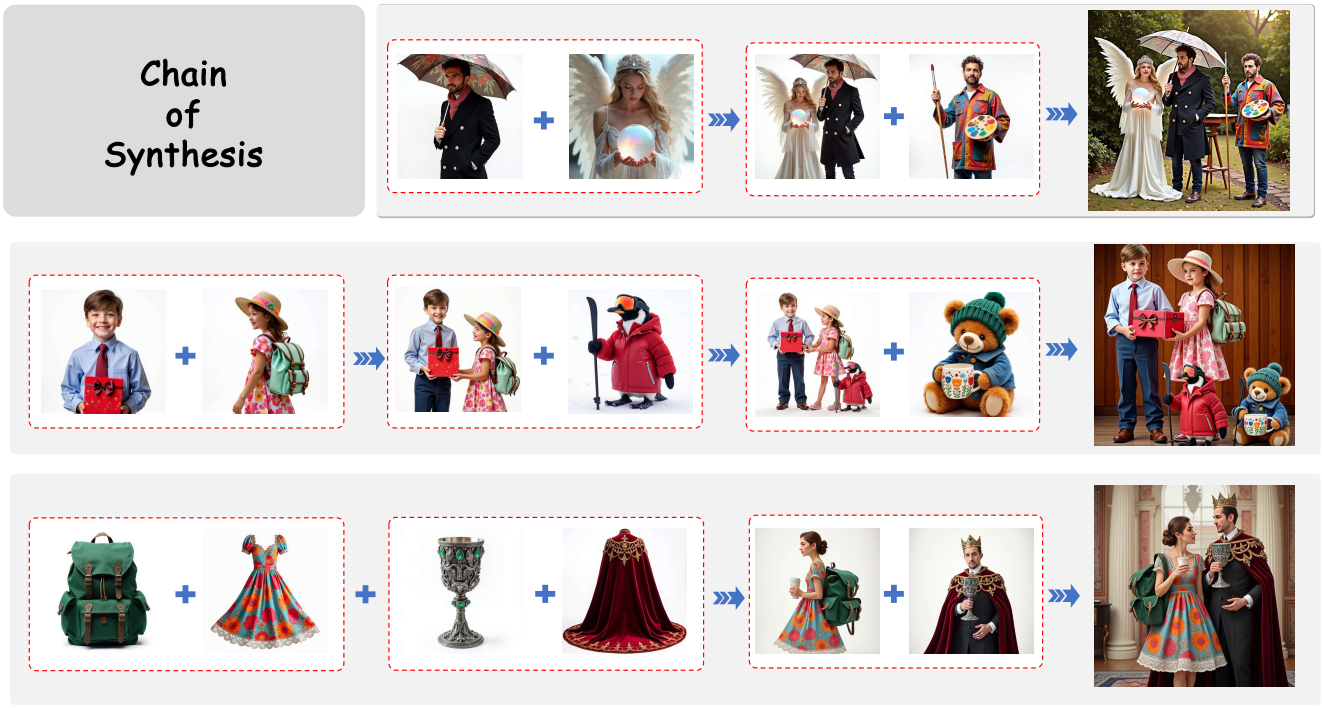

Figure 1. The intermediate results of the Chain of Synthesis.

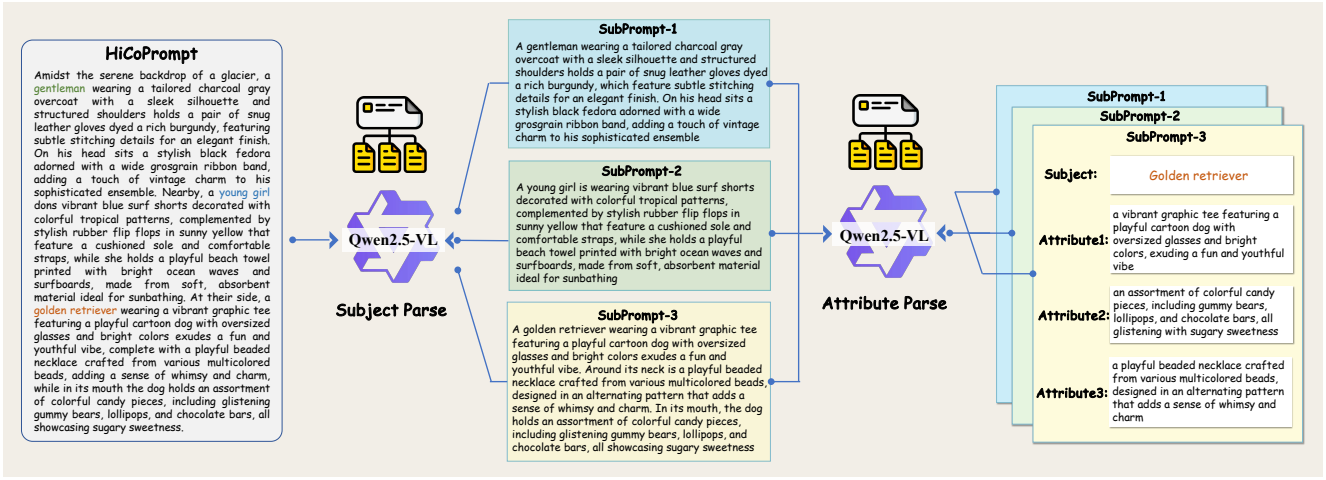

Figure 2. The intermediate prompt of the Chain of Synthesis (zoom in for details).

1. Implementation details

The weights of the reward are set to $w_{\text{clip}} = 0.7, w_{\text{hps}} = 1.4, w_{\text{dino}} = 0.7, w_{\text{vlm}} = 0.7$. The size of the trainset and testset is set to 12,000 and 3,000, respectively. The number of de-noise steps is set to 16 when sampling. The $\eta_{\max}$ and $\eta_{\min}$ are set to 1.0 and 0. The output image size is set to 512×512, with the reference image size being 320×320. The total training step is set to 300 with a batch size of 12.

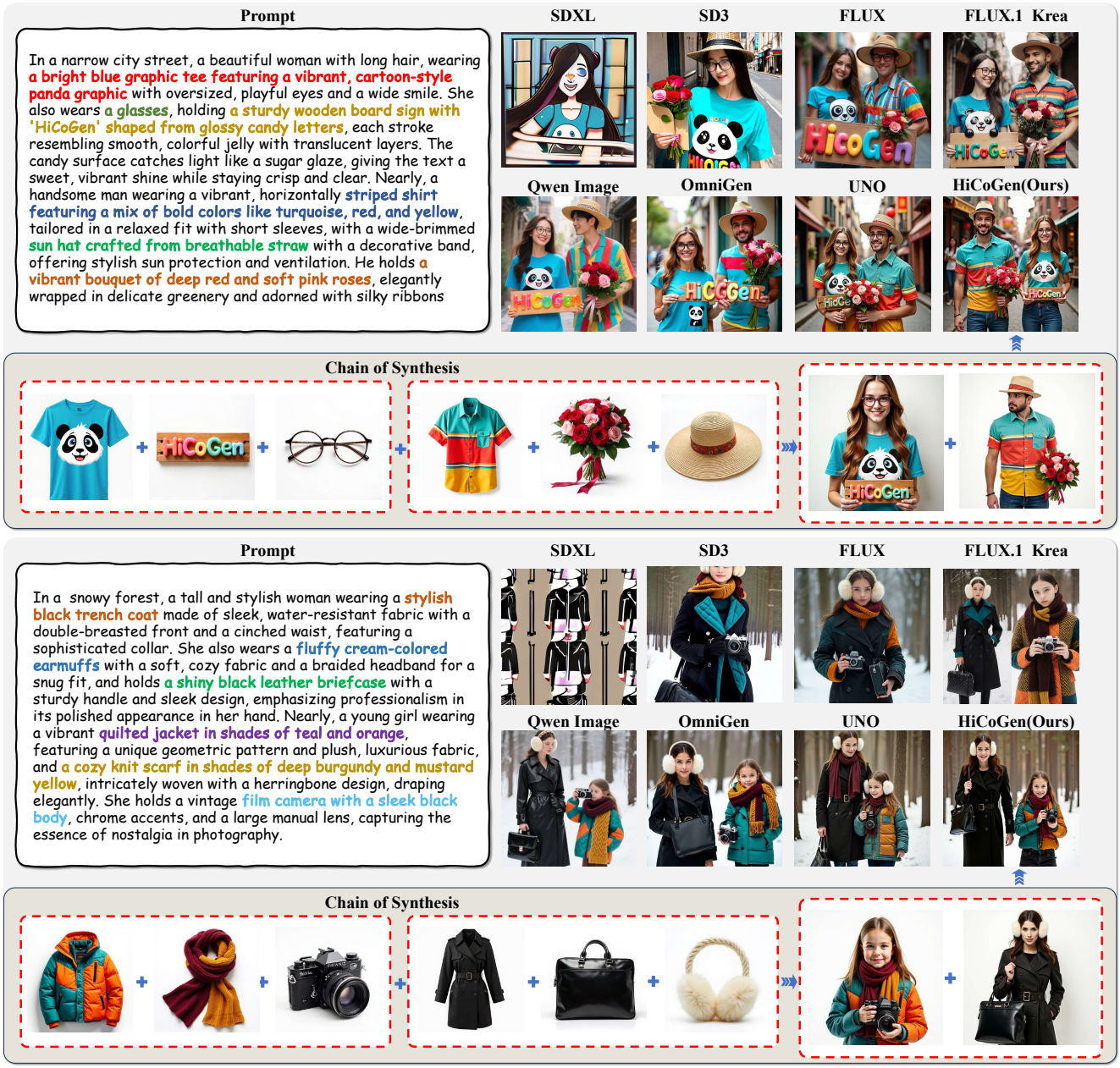

Figure 3. More visual results of the HiCoGen

2. Chain of Synthesis

Fig. 1 shows the case of how the HiCoGen performs Chain of Synthesis. The first row showcases the CoS of three subjects in HiCoGen. It first generates two subjects in one intermediate output, and then includes the third subject in the final output. Besides, we also present the scaling results of four subjects in the second row. The third row shows the result of the HiCoGen on how to generate subjects with hierarchical relationships. For example, it generates clothing and items of the subject, and then generates the final image. Fig. 2 indicates the whole workflow of the parse and rewrite LLM. For a long and complex HiCoPrompt, it is first parsed according to subjects (e.g., gentleman, young girl, and golden retriver). Then, the sub-prompt is continued to be parsed if it contains multiple attributes that can be generated. Thus, we replace the long and complex prompt with a structured prompt composed of several semantic units.

3. More Visual Results

Fig. 3 presents additional visual results produced by HiCoGen, as well as the results from other methods. We also illustrate the CoS process involved in the generation. Through CoS, HiCoGen generates all the concepts from the HiCoPrompt.

Subject Generation

Role:
Please be very creative and generate 20 groups of components for the given character.

Follow these rules:
1. You will be given a <character>, you need to create its wearing, holding, and accessory items.
2. These items must exist in the real world (no fantasy or fictional materials).
3. Do not repeat the same words between outputs; use your imagination and common sense of real life.
4. Each item must contain **no more than two words**.
5. Output multiple unique sets if possible.
6. Given a final complete and brief prompt for text-to-image generation

Output format:
[character]: <character name>
[clothing1]: <real-world clothing>
[holding1]: <real-world object>
[accessory1]: <real-world accessory>

Example:
[character]: dog
[clothing1]: superman's costume
[holding1]: sign
[accessory1]: goggles
[brief_prompt1]: a dog wearing a superman's costume with a goggles, holding a sign

[clothing2]: space suit
[holding2]: book
[accessory2]: glasses
[brief_prompt2]: a dog wearing a space suit with glasses, holding a book
...
(Up to [asset20])

[character]: character

Attribute Rewrite

Role:
Please be very creative and generate a detailed prompt for text-to-image generation.

Follow these rules:
1. You will be given 3 {assets}, you need to create an asset (detailed subject prompt) based on the {assets}.
2. These descriptions can refer only to appearance descriptions/or to certain brands. e.g., "Elon Musk in pajamas", "a tiger in a black hat", "A Mercedes sports car", "A blonde", "A rotten wooden door", "a book with cover written 'Magic'"
3. Do not repeat the same words between outputs; use your imagination and common sense of real life.
4. Describe this asset in one sentence. No more than 40 words
5. Focus on the asset itself, and must contain the asset in each output.
6. Do not describe its background and environment.
7. You may rewrite it based on its contextual meaning within the original prompt.
8. Do not include any users, wearers, or owners. Describe only the object itself.

Example:
[input_asset1]: book
[input_asset2]: lab coat
[input_asset3]: necklace
[original_prompt]: a beautiful woman wearing a lab coat with a necklace and holding a book.

[output_asset1]: an open ancient magic book with a thick dark brown tanned leather cover, adorned with hand-embossed golden runes and intricate patterns
[output_asset2]: a crisp white lab coat with embroidered name and pen-stained pocket
[output_asset3]: a whimsical necklace with animal-shaped pendants

Now, please generate the detailed prompt for the following assets:
[input_asset1]: {subject1}
[input_asset2]: {subject2}
[input_asset3]: {subject3}
[original_prompt]: {ori_prompt}

Prompt Reframing

Role:
You are a professional prompt structure optimization and synthesis expert, skilled at merging multiple prompt segments into a single, well-structured and logically coherent prompt.

Follow these rules:
Generate a unified prompt based on the provided ori_prompt (original prompt) and subprompts (refined sub-prompts). Please strictly follow these rules:
1. Preserve the internal structure, hierarchy, and semantics of each subprompt.
2. You may adjust sentence structures or logical order slightly during merging to ensure overall fluency and coherence.
3. Do not delete or rewrite any key information from the subprompts.
4. You may adapt the language style according to the tone or purpose of the ori_prompt (e.g., analysis, creation, reasoning, Q&A, etc.).
5. The final output should clearly reflect the role and integrated content of each subpart.
6. Output one complete prompt paragraph — no explanations or additional commentary are needed.
7. Do not describe its background and environment.

[input_asset1]: {subject1}
[input_asset2]: {subject2}
[input_asset3]: {subject3}
[original_prompt]: {ori_prompt}

4. Prompt

4.1. Dataset Creation Prompt

In this section, we demonstrate the prompt used to construct the dataset. First, a subject is randomly selected from a predefined character pool. Then, this subject will be assigned up to three attributes. After that, these attributes are rewritten into a more detailed version. Finally, these prompts are reframed into the final prompts by LLMs.

Subject Reward

Role
You are an expert AI assistant specializing in the objective evaluation of the consistency of subjects in two images. Assign a specific integer score to subject. More and larger differences result in a lower score.

Important Notes
- Provide quantitative differences in reason whenever possible.
- Ignore differences in the subject's background, environment, position, size, etc.
- Ignore differences in the subject's actions, poses, expressions, viewpoints, additional accessories, etc.
- Ignore the extra accessory of the subject in the second image, such as hat, glasses, etc.

You must adhere to the output format strictly.

The score ranges from 0 to 4:
- 0: No resemblance. The {subject} in Image2 does NOT appear in Image1 at all. No matching identifiable features.
- 1: Minimal resemblance. The {subject} in Image2 appears only minimally in the {subject} in Image1. Only trivial or coincidental similarities.
- 2: Moderate resemblance. The {subject} in Image2 partially matches the {subject} in Image1. Some major features match, but key identity traits differ.
- 3: Strong resemblance. The {subject} in Image2 is very similar to the {subject} in Image1. Most identity traits align, with minor differences.
- 4: Near-identical. The {subject} in Image2 is identical to the {subject} in Image1. Nearly identical; same character.

Output only the brief reason and score. DO NOT OUTPUT any other text.

Image1: [brief description of Image1]
Image2: [brief description of Image2]
Reason: [Brief Reason]
Output: [Score]

Relation Reward

Role:
You are a visual consistency evaluator. You will compare Image2 and Image1.

Your task is to check two things:
1) Content Preservation:
Are the key visual elements from Image2 present in Image1? Ignore differences in the subject's background, environment, position, size, etc.
Examples of key visual elements:
- For a character: clothing type, accessories.
- For clothing: color, pattern, shape, length, structure.
- For a held object: object type, shape
2) Correct Assignment:
The elements of the character from Image2 must appear on the same character in Image1.
They must not be assigned to another character or misplaced.

Then classify overall similarity using the scale below:

0 - Completely mismatched: Key elements are missing or assigned to the wrong subject. No meaningful match.
1 - Minimal similarity: Only small or generic resemblance. Most defining elements are missing or incorrectly reassigned.
2 - Partial similarity: Some important features match, but major elements are missing, altered, or incorrectly assigned.
3 - High similarity: Most key features are present and assigned correctly, with only minor inaccuracies.
4 - Fully consistent: All key features are preserved and assigned correctly. No confusion or mixing.

Output only the following format:

Image1: [brief description of Image1]
Image2: [brief description of Image2]
Reason: [brief explanation of match or mismatch]
Output: [0-4]

4.2. Reward Prompt

Accuracy of Existing

Check whether a specific object appears in the image.

Instruction:
Analyze the provided image carefully. Determine whether the specified item is present or visible in the image.

Object: {user_prompt}

Output format:
Answer "Yes" if the object is clearly present.

Answer "No" if the object is not visible. Do not assume or infer its presence based on context.

If uncertain, answer "Unclear" and briefly explain why.

Output only the brief description, reason and answer. DO NOT OUTPUT other text.

Description: [Brief Description of the Image]
Reason: [Brief Reason]
Output: [Yes / No / Unclear]

Accuracy of Attribute

Check whether the object in the image matches the attributes described in the given text.

Instruction:
Analyze the image carefully. Determine whether the object matches the specific details (such as color, shape, material, pattern, size, part structure, or other explicit attributes) described in the text. Object description: "{user_prompt}"

Output format:

Answer "Yes" if the object is clearly present and its visible attributes match the given text.

Answer "No" if the object is present but clearly does not match the described details, or the described object is not visible at all.

If uncertain, answer "Unclear" and briefly explain why.

Description: [Brief Description of the Image]
Reason: [Brief Reason]
Output: [Yes / No / Unclear]

Accuracy of Attribute

Check whether the specified object in the image interacts with the main subject in a correct and reasonable way according to the following rules.

Rules:
The object mentioned in the prompt must interact with the main subject in a reasonable, physically consistent way (no floating, no paste-on appearance).
The object must not appear in the hands or possession of another subject.
The object's size and proportion relative to the main subject must be correct and not distorted.

Instruction:
Analyze the provided image carefully and judge whether the described object satisfies all the above rules.
Object: "{user_prompt}"

Output format:
Output "Yes" if the object is visible and satisfies all three rules.
Output "No" if any rule is violated or the object is not visible.

Description: [Brief Description of the Image]
Reason: [Brief Reason]
Output: [Yes / No / Unclear]

4.3. Evaluation Prompt