
Vinci: Deep Thinking in Text-to-Image Generation using Unified Model with Reinforcement Learning

Anonymous Author(s)
Affiliation
Address
email

Contents

A Dataset Construction	1
A.1 MCoT Construction Pipeline	1
A.2 Distribution of Reasoning Text Lengths	1
B QA-based Reward for Generated Text	2
C Implementation Details	3
C.1 Training Details	3
C.2 Data Preparation	3
D Societal Impacts	3

A Dataset Construction

A.1 MCoT Construction Pipeline

To construct our Multimodal-CoT(MCoT) dataset, we designed a multi-stage pipeline comprising Image Generation for MCoT, Text Generation for MCoT, and MCoT Construction. Given a text query q , we first generated a set of n candidate images $\{p_1, p_2, \dots, p_n\}$ using FLUX.1-dev [2]. To identify objects, we applied Mask2Former [1] to each image, yielding detection outputs o_i and forming the set of tuples $\{(p_1, o_1), (p_2, o_2), \dots, (p_n, o_n)\}$, where o_i denoted the objects detected in image p_i . Then, to mitigate ambiguity in multi-image interpretation, we tasked Qwen-VL [3] with independently evaluating each tuple (p_i, o_i) for its semantic alignment with the original query and overall quality. The detection result o_i was explicitly provided to the model to reduce hallucinations. During this evaluation, the model produced a caption c_i and a quality score g_i , resulting in annotated triples (p_i, c_i, g_i) . These were sorted in ascending order by g_i to yield a ranked sequence of annotated samples. Finally, to construct the MCoT data, we randomly selected k triples from the sorted list while preserving their relative score order. Each selected set must contain at least one high-quality (i.e., high-scoring) triple, which served as the final reasoning step. The resulting combination $\{(p_i^1, c_i^1, g_i^1), \dots, (p_i^k, c_i^k, g_i^k)\}$ was input into Qwen-VL to construct the final MCoT data. The prompts used by Qwen-VL during the construction process are shown in Table 2.

The final MCoT dataset, consisting of 20,000 samples, comprises three types: 1-step MCoT, 2-step MCoT, and 3-step MCoT, which correspond to reasoning processes involving 1, 2, and 3 steps before generating the final image, respectively. Examples of these three types are shown in Table 1.

A.2 Distribution of Reasoning Text Lengths

A key characteristic of MCoT data is the ability to clearly understand and articulate the content of images generated throughout the reasoning process. Therefore, the length of the accompanying text is

MCoT Data Type	Example
1-step MCoT	<p><think></p> <p>The user asked to generate an image: a photo of two tennis rackets. One possible image was <1_image>. The image clearly reflected the prompt, showing two rackets placed side by side with distinct yet harmonious designs. The rackets were rendered in sharp detail, with red and black colors creating a vivid contrast. A soft, neutral background kept focus on the objects. Lighting and composition gave the image a polished, professional look, making it a faithful and visually appealing realization of the prompt. Overall, <1_image> was an excellent realization of the prompt, and I selected it as answer because it not only met the basic requirements but also elevated the concept through its high-quality execution and attention to detail.</p> <p></think></p> <p><answer><1_image></answer></p>
2-step MCoT	<p><think></p> <p>The user asked to generate an image: a photo of a sandwich below a sports ball. One possible image was <1_image>. While <1_image> did present a sandwich with a sports ball above it, the sandwich appeared somewhat simplistic and lacked detailed texturing. I needed to generate an image that clearly showed a sandwich below a sports ball, with the sandwich rendered in a more realistic and aesthetically pleasing way.</p> <p>Then, I generated <2_image>, which offered a more refined and visually compelling representation of the prompt. The image clearly showed a sports ball above a sandwich, and their spatial relationship was accurate. The sandwich in <2_image> had a more realistic texture, with a visibly toasted bun that added depth and detail. The layers, including lettuce and other fillings, were vibrant and well-defined.</p> <p>In summary, <2_image> correctly depicted a sandwich below a sports ball with a clear spatial arrangement. Both the sandwich and the sports ball featured improved textures and fine-grained details, contributing to the overall quality. These factors made <2_image> the preferred choice over <1_image>.</p> <p></think></p> <p><answer><2_image></answer></p>
3-step MCoT	<p><think></p> <p>The user asked to generate an image: a photo of an orange backpack and a green spoon. One possible image was <1_image>. However, it did not include an orange backpack as expected, which was a critical element of the prompt. I needed to generate an image that included an orange backpack.</p> <p>I then generated <2_image>, which included the required orange backpack and aligned better with the prompt. While <2_image> was an improvement, its color scheme was relatively dull, and the background did not effectively complement the orange backpack. I needed to make the colors more vibrant.</p> <p>Finally, I generated <3_image>, which featured a more vibrant and consistent color palette. The orange backpack and green spoon stood out clearly against a light green background. The lighting was evenly distributed, enhancing the textures and colors of the objects, which appeared more realistic and detailed.</p> <p>In summary, I selected <3_image> as the final result because it accurately captured both the orange backpack and the green spoon with a clear and visually pleasing composition.</p> <p></think></p> <p><answer><3_image></answer></p>

Table 1: Introduction of the three types of MCoT data used in our cold-start stage. <1_image>, <2_image>, and <3_image> represent images generated after 1, 2, and 3 rounds of reasoning, respectively. The corresponding visual tokens are omitted for brevity.

33 an important aspect to consider. The distribution of reasoning text lengths in our constructed MCoT
34 data is shown in Figure 1.

35 B QA-based Reward for Generated Text

36 In our research, we designed a QA-based reward function to evaluate the quality of the generated text
37 in the context of text-to-image generation. This reward function assesses whether the generated text
38 accurately describes the generated image, identifies any issues in the image, and provides strategies
39 for improvement. In Table 3, we provide examples of the prompts used to guide the evaluation
40 process.

41 In our implementation, we used a multi-modal large language model (MLLM) to evaluate the
42 generated text based on the prompts provided above. The scores and explanations provided by the
43 MLLM were then used as part of the overall reward function to guide the training of our text-to-image
44 generation model. By incorporating this QA-based reward function, we aimed to enhance the model’s
45 ability to generate high-quality, contextually accurate, and self-reflective image descriptions, thereby
46 improving the overall performance of the text-to-image generation process.

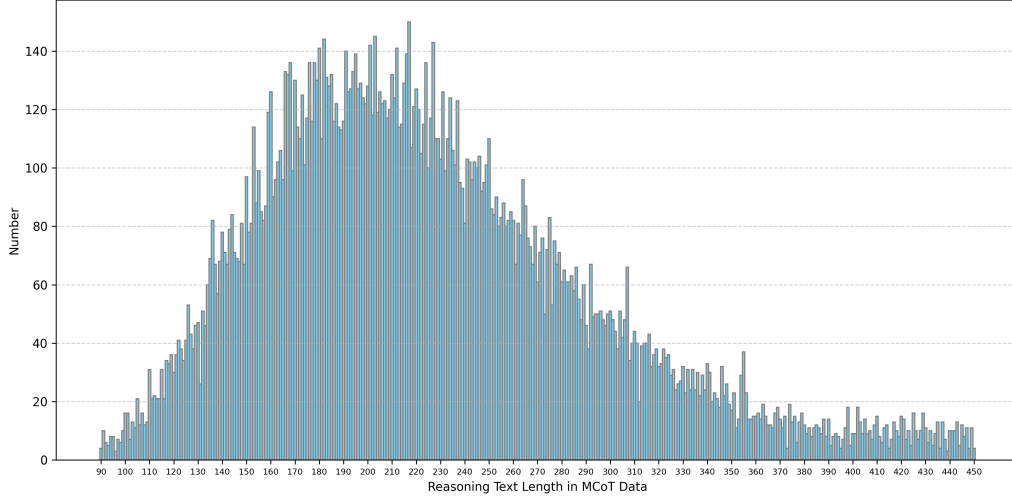


Figure 1: Distribution of reasoning text lengths in our MCoT dataset

47 C Implementation Details

48 C.1 Training Details

49 For our training, we adopted Emu3-Gen [4] as the base model, which has a unified image representa-
 50 tion. The model’s context length was extended to 15,360 tokens, which allowed for processing longer
 51 sequences of text and image pairs and supported approximately three iterations within each context
 52 window. The learning rate was set to $1e-5$ and a beta of 0.01.

53 During the training, we utilized 16 A800 GPUs. The training was divided into two stages. In the first
 54 stage, known as the cold start, we set the batch size to 64 and trained for approximately 20 hours,
 55 allowing the model to learn the fundamental features and patterns of the data. In the second stage,
 56 which involved reinforcement learning, we configured the group size to 8 and trained for 60 hours,
 57 further optimizing the model’s performance through external feedback mechanisms.

58 C.2 Data Preparation

59 All of our images are in the resolution of 512×512 , and vision tokens were generated using vision
 60 tokenizer of Emu3. Following the Emu3 design, we incorporated five special tokens to merge textual
 61 and visual data, constructing interleaved vision-language MCoT data and document-like inputs for
 62 the training process.

63 Taking the 2-step MCoT data as an example, whenever `<1_image>` or `<2_image>` appears for the
 64 first time in the MCoT sequence, it is immediately followed by the corresponding image token block.
 65 The resulting training data can be structured as follows:

66 `<1_image>[BOI]{meta text}[SOT]{vision tokens}[EOL][EOF][EOI]`

67 Here, the image token block begins with `[BOI]`, where `{meta text}` contains information about the
 68 image resolution. The token `[SOT]` marks the beginning of the vision token sequence. Additionally,
 69 `[EOL]` and `[EOF]` are inserted into the token stream to indicate line breaks. The image token block
 70 ends with `[EOI]`.

71 D Societal Impacts

72 Vinci, as a novel framework that integrates deep reasoning capabilities into text-to-image generation
 73 through a unified model with reinforcement learning, has the potential to significantly enhance
 74 various applications that rely on visual content creation. One of the key positive impacts is the
 75 improvement in the quality and alignment of generated images with textual prompts. This capability
 76 can greatly benefit educational and creative industries. For instance, in education, Vinci can generate

Stage	Prompt Example
Text Generation	<p>You are tasked with generating a caption for a given image and its detection results, and evaluating how well the image aligns with the original prompt in terms of semantic accuracy and generation quality.</p> <p>Your task involves the following steps:</p> <ol style="list-style-type: none"> 1. Caption Generation <ul style="list-style-type: none"> - Generate a coherent and informative caption that accurately describes the given image, using the provided object detection results as a reference. - The caption should cover key visual elements, including object types, positions, colors, and spatial relationships, and reflect the intent of the original prompt. 2. Prompt-Image Alignment and Generation Quality Evaluation <ul style="list-style-type: none"> - Evaluate how well the image itself matches the original prompt in terms of semantic content. - In addition, assess the visual quality of the generated image, including realism, clarity, and overall aesthetic quality. 3. Scoring <ul style="list-style-type: none"> - Provide an overall score on a scale from 0 to 4 based on both prompt alignment and image quality: - 0 indicates the image does not align with the prompt and is of low visual quality. It fails both semantically and aesthetically. - 1 indicates the image is visually decent but fails to capture the core semantics of the prompt. It may contain hallucinated or unrelated content. - 2 indicates the image is partially aligned with the prompt and has moderate quality. Some key elements may be missing or inaccurately rendered. - 3 indicates the image correctly reflects the prompt but suffers from low visual quality (e.g., blurry, distorted, or unnatural rendering). - 4 indicates the image is fully aligned with the prompt and of high visual quality. It accurately presents all required elements in a realistic, clear, and aesthetically pleasing manner.
MCoT Construction	<p>You are given a sequence of (image, caption, score) triples ranked in ascending order by their score and their original prompt, where each triple consists of:</p> <ul style="list-style-type: none"> - an image generated based on a text prompt, - a caption describing the image, - a score indicating how well the image aligns with the original prompt in terms of semantic relevance and generation quality. <p>Your task is to simulate a step-by-step reasoning process that leads to the final decision about which image best satisfies the original prompt. This process should reflect how a human might evaluate and revise image generations based on feedback and visual inspection.</p> <p>Please proceed as follows:</p> <ol style="list-style-type: none"> 1. Analyze each image and caption in order, reflecting on what aspects are missing, incorrect, or can be improved. 2. Describe how the reasoning evolves across steps and why one image is better than the previous ones. 3. End the reasoning by selecting the best image and briefly summarizing why it is the final choice. <p>Output:</p> <p>Your response must strictly follow the format below:</p> <pre> <think> The user asked/requested to generate an image: [prompt]. One possible image is <1_image> [your generation and reasoning process] </think> <answer>[your choice]</answer> </pre>

Table 2: Introduction of prompts used in MCoT construction

highly accurate and contextually relevant illustrations for textbooks, making learning materials more engaging and accessible for students. In creative industries, such as graphic design and advertising, Vinci can assist designers in quickly generating high-quality visual concepts, thereby accelerating the creative process and potentially leading to more innovative and diverse visual content.

Despite its potential benefits, Vinci also poses several risks that need to be carefully considered. One of the primary concerns is the potential misuse of the technology for generating misleading or harmful visual content. Vinci’s ability to generate high-quality images based on textual prompts increases the risk of creating deepfakes or manipulated images that could be used to spread disinformation, manipulate public opinion, or harm individuals’ reputations. The ease with which these images can be generated and disseminated poses a significant threat to societal trust and information integrity.

Example

You are tasked with evaluating the quality of an image description generated by a text-to-image model. Please provide a detailed evaluation based on the following aspects:

1. Completeness of the Image Description:

- Assess whether the description covers all the key elements and details present in the image. Consider whether it includes descriptions of objects, their positions, colors, and any other relevant visual attributes.

2. Identification of Issues in the Image:

- Determine if the description identifies any discrepancies or issues in the generated image compared to the original prompt. This could include missing elements, incorrect colors, misplaced objects, or any other inconsistencies.

3. Strategies for Improvement:

- Suggest specific strategies or adjustments that could improve the accuracy and quality of the generated image. This could involve changes to the prompt or other recommendations to enhance the alignment between the generated image and the original prompt.

Image Description: [Generated Image Description]

Please provide a score for each aspect on a scale of 0 to 2, where:

- 0 indicates poor performance,
- 1 indicates average performance,
- 2 indicates excellent performance.

Additionally, provide a brief explanation for each score to justify your evaluation.

Table 3: Introduction of the prompt used to evaluate the generated text's quality.

87 **References**

- 88 [1] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar.
89 Masked-attention mask transformer for universal image segmentation. In *Proceedings of the*
90 *IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022.
- 91 [2] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024.
- 92 [3] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu,
93 Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng
94 Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language
95 model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.
- 96 [4] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan
97 Zhang, Yueze Wang, Zhen Li, Qiyang Yu, et al. Emu3: Next-token prediction is all you need.
98 *arXiv preprint arXiv:2409.18869*, 2024.

99