Figure 7: **More qualitative results across different resolutions.**382 **A.1 Implemnetary Details**

383 In the first stage of training, we utilize the Laion COCO [30] and JourneyDB [33] image-caption pairs
 384 as foundational text2image data. Additionally, we incorporate the X2I dataset [45], which includes
 385 tasks for computer vision, in-context learning, multi-modal instructions, and subject-driven. To
 386 further enrich our dataset, we leverage Sana [46] and Flux [18] to generate high-quality images based
 387 on the existing prompt. During training, we use a constant learning rate scheduler with an initial
 388 rate of 1×10^{-4} and a weight decay of 0.05. The model is trained with a batch size of 1024 using

images at a resolution of 256×256 . In the second stage, we employ Qwen2.5-VL [1] to generate CoT instruction data. At this stage, we progressively increase the training resolution to 512×512 , while reducing the learning rate to 5×10^{-5} . In the final stage, we utilize Qwen3 [38] to generate a dataset of logical reasoning texts, which serve as training data for reinforcement learning. We adopt a cosine scheduler during the reinforcement learning phase.

We evaluate our method during multimodal understanding and various generation benchmarks. Specifically, our model is evaluated on MMMU [51], MMBench [50] and RealworldQA for image understanding. As for basic image generation tasks, we evaluate our model on GenEval [10], which involves various metrics such as counting, colors, and position. We also evaluate our text-to-image generation capability on DPG-Bench [14] following previous methods [4, 28]. For reasoning generation evaluation on WISE [21] Benchmark, we perform our RG-VLM into thinking mode and generate the corresponding image with 1024×1024 .

A.2 Detailed Response of Our RG-VLM

For the user input of different reasoning generations in Fig. 1, the detailed response of our model RG-VLM is shown in Fig. 8, Fig. 9, Fig. 10 and Fig. 11.

User: Generate an image of an animal with $(3 + 6)$ lives.

RG-VLM: **<think>** To generate an image based on the instruction "An image of an animal with $(3 + 6)$ lives," we first need to interpret the mathematical expression given: $(3 + 6 = 9)$. This suggests that the animal should have nine lives, which is a common characteristic associated with mythical creatures like cats or dragons.

Given this interpretation, I would suggest creating an image of a cat, which is known for having nine lives. The image could be a digital drawing or illustration of a cat, possibly with a playful and whimsical style to reflect its enduring and resilient nature.

The instruction does not specify the type of animal further, but commonly, a cat is used when referring to the concept of "nine lives" due to the popular belief that a cat can survive falls from great heights, often attributed to its ability to land on its feet.

For the image:

- The cat could be depicted in a simple yet vibrant drawing.
- It might be shown in a dynamic pose, perhaps mid-leap or after a sudden turn.
- The cat could have a light, fluffy appearance with large eyes and a small body, capturing the classic cartoonish representation of a cat.

This approach aligns with the whimsical and imaginative nature typically associated with the idea of a cat having nine lives. **</think>**

<answer> To create an image according to the instructions, you could generate a digital drawing or illustration of a cat. The cat should have a playful and whimsical style, possibly depicted in a dynamic pose such as mid-leap or after a sudden turn.

The cat could be drawn with a light, fluffy appearance, large eyes, and a small body, capturing the classic cartoonish representation of a cat. This image would visually represent the concept of the cat having nine lives, as traditionally associated with feline resilience and agility. ****

</answer>



User: An elephant and a rabbit stand on both sides of a seesaw.

RG-VLM: **<think>** To generate an illustration of an elephant and a rabbit standing on both sides of a seesaw, we need to consider a few key elements: the size difference between the elephant and the rabbit, the balance and movement implied by the seesaw, and the positioning of the animals. Elephants are much larger than rabbits, so they would be on one end of the seesaw, while the rabbit would likely be at the other, opposite end.

The seesaw is typically represented as a horizontal bar balanced on a central fulcrum. The elephants and the rabbit would each be standing on a platform or platform-like structure that extends from either side of this central bar. **The elephant's end would be lower because its mass is significantly greater than the rabbit's. The rabbit's end would likely be higher due to the balancing effect of the seesaw.**

In terms of style, the illustration could be cartoonish for a simple, child-friendly look, or more realistic if aiming for a detailed and lifelike depiction. The background can be minimal, perhaps just suggesting an open space, which keeps the focus on the main subjects.

To create such an image, one might start with a simple sketch, then add details like the elephant's trunk, the rabbit's long ears, the seesaw's balance point, and the platforms. Color choices could include earthy tones for the elephant (like brown) and lighter, more playful colors for the rabbit (like pink or yellow). **</think>**

<answer> Creating an image of an elephant and a rabbit on a seesaw involves drawing or digitally rendering two large elephants and a small rabbit placed at opposite ends of a seesaw. The elephant should be standing on a platform near one end of the seesaw, and the rabbit should be on the opposite end. The seesaw should be shown balancing in the middle point, with the elephant's end appearing lower due to its larger mass and the rabbit's end appearing higher.

For a simple illustration:

1. Sketch the outline of the seesaw with the fulcrum in the center.
2. Add the elephant on one side with its trunk possibly hanging down or interacting, and the rabbit on the other side.
3. Color the elephant in darker shades, the rabbit in lighter, playful colors.
4. Ensure the seesaw is balanced correctly, with the elephant's end lower and the rabbit's end. ****

</answer>



Figure 8: Detailed response of our RG-VLM.

403

A.3 Limitation

In this work, our Chain-of-Thought (CoT) content is presented in plain text, while multimodal CoT processes will be addressed in future research.

A.4 More Qualitative Results

We provide more qualitative results about text2image generation of our RG-VLM as shown in Fig. 7.

User: A pool at minus ten degrees Celsius.

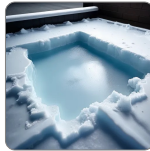
RG-VLM : <think> To generate an image of a pool at minus ten degrees Celsius, one should consider the freezing point of water (0°C or 32°F) and the consequences of the water freezing below this temperature. At minus ten degrees Celsius, water would indeed be frozen into ice, forming a solid surface. The appearance would likely include a clear, uncolored ice layer covering the pool's water, which, due to the cold, would have a glassy, possibly cloudy texture. The image would depict a scene where the entire body of water is frozen, creating a hard, icy surface on top.

To create such an image, it might involve drawing or digitally rendering a rectangular or circular pool shape filled with a smooth, flat, transparent ice layer. The background could show the outline of the pool with visible edges, suggesting a boundary for the frozen area. The ice surface can be depicted in a light blue or white color, with subtle shading to indicate the thickness of the ice and any slight irregularities that could be present in real-life scenarios. The image should not include any liquid water, as at minus ten degrees Celsius, the water has already frozen.

The visual representation will be static, as the image would not capture movement; it would focus solely on the frozen state of the water, highlighting the transformation from liquid to solid under extreme cold conditions.</think>

<answer> To create the image, start by drawing a rectangular or circular outline representing the pool. Fill the pool with a flat, transparent ice layer that covers the entire surface. The ice can be rendered in a light blue or white color, with slight variations in hue to suggest thickness and the potential for small air bubbles or impurities trapped in the ice. Ensure the ice layer is smooth but shows some variation in texture to reflect its real-world appearance.

The overall image should emphasize the immobility and stillness of the frozen water, with no indication of liquid water beneath the ice. The icy surface should be depicted as a hard, solid cover over the entire pool area, with no thawing or melting visible in the image. The environment surrounding the pool should be kept neutral to avoid drawing attention away from the main subject—the frozen pool. </answer>



User: Scene in the Sydney Opera House when New York is at noon.

RG-VLM : <think> To generate an image of the Sydney Opera House when New York is at noon, we need to consider the geographical and time difference between Sydney (which is in the Eastern Australia Time Zone) and New York (which is in Eastern Standard Time). The time difference between these two locations is approximately 15 hours, with New York being 15 hours ahead of Sydney. Therefore, when it's noon in New York, it would be approximately 3 AM in Sydney.

The image should reflect the interior of the Sydney Opera House during this late hour of the night or early morning. The distinctive curved shell-like structures characteristic of the Opera House would likely be illuminated by the artificial lighting inside the building, which would contrast with the dark surroundings outside. The seating areas, corridors, and architectural features would be highlighted.

Given that the specific time would be close to dawn, there might be a subtle hint of light or the appearance of sunrise through windows or doors leading to the interior.

In terms of capturing this scene artistically, the focus would be on the interplay of shadows and light within the interior design of the Opera House, emphasizing the futuristic and modern aesthetic of the structure. The overall atmosphere would be calm and serene, with the darkness outside juxtaposed against the bright interior lights.</think>

<answer> This could be best conveyed through a digital illustration or a photo composition. You might use software like Adobe Photoshop or digital drawing tools where you draw or composite elements of the interior of the Sydney Opera House, specifically highlighting its iconic architectural features like the white curves and the rows of seating, all rendered inside the building.

The scene would be set in a still, minimalist environment mimicking 3 AM, with ambient lighting creating soft patterns across the surfaces of the seats and the ceiling. You could add a faint glow from the windows showing the beginning of daylight outside to provide a sense of the time. The lighting within the Opera House would depict a variety of light sources: warm accent lights for dramatic effect, and the soft, scattered lights from a dim evening ambience. </answer>



Figure 9: Detailed response of our RG-VLM.

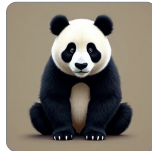
User: An image of China's national treasure animal.

RG-VLM : <think> To generate an image of China's national treasure animal, one should consider which animal is commonly recognized as the "national treasure" in Chinese culture. This is often the giant panda (Ailuropoda melanoleuca). Pandas are iconic symbols of China and are known for their distinctive black-and-white fur pattern.

The instruction does not specify the medium or style of the image, so we can create a digital illustration of a panda. Pandas typically have a round face, large dark eyes, and a body that is mostly white with black markings around their eyes, ears, legs, and shoulders.

For the style, a simple cartoon-like or realistic style could be used, depending on the intended audience and context. If the image is meant to be educational or for children, a cartoonish style might be more appropriate. For a more formal or artistic representation, a more detailed and realistic depiction would be suitable.</think>

<answer> To create an image of China's national treasure animal, such as a giant panda, you could start by drawing the characteristic features: a round face, large dark eyes, and a body that is predominantly white with black patterns around the eyes, ears, legs, and shoulders. A simple cartoon style or a more detailed, realistic rendering can both capture the essence of a panda. </answer>



User: An image of multiple apples, the quantity of apples is the solution to the equation " $x^2 + 2 = 11$ ".

RG-VLM : <think> To generate the image, we need to solve the given equation for x first: $(x^2 + 2 = 11)$. Subtracting 2 from both sides gives us $(x^2 = 9)$. Taking the square root of both sides, we get $(x = \pm 3)$. Since the number of apples must be a positive integer, we consider only the positive solution, which is 3. Therefore, the image should depict three apples.

The next step is to create an image that visually represents this solution. This could be a simple drawing or a digital illustration showing three apples. The apples can be depicted in any style, such as a cartoonish or realistic representation, but the key point is that there should be exactly three apples in the image.</think>

<answer> To create the image, you would draw or digitally render three apples. The apples can be placed in a row, stacked, or arranged in any other way that clearly shows the number three. The image should be simple enough to convey the solution to the equation $(x^2 + 2 = 11)$ by visually representing the quantity of apples, which is 3.

You might use a white background with the apples in a natural color like red or green, depending on your preference. If you're using a digital tool, you can draw the apples directly or use clip art to place them in the image. The focus should be on clarity and simplicity to ensure the viewer can easily understand that the image represents the solution to the equation, which is three apples. </answer>



Figure 10: Detailed response of our RG-VLM.

User: <IMG1> Generate an image of the animal growing up.



RG-VLM : <think> To generate an image according to the given instruction, we need to consider what the tadpole will look like once it has grown into a frog. **Typically, a frog is characterized by its four legs, a flat body, and a prominent head with two eyes on top. The tail, which is a distinct feature of the tadpole stage, will have disappeared as the frog develops.** The transition from tadpole to frog involves significant physical changes that include the development of hind legs before the front legs,

the growth of the lungs for breathing air, and the transformation of the tail into hind legs. The skin also thickens, becoming smoother and drier.

Given the instruction to generate an image of the animal after growing up, I will create a depiction of a fully developed frog, which is the adult form of the frog species depicted here.

</think><answer> To generate an image of a frog after the tadpole stage, you might imagine a frog with a robust body, smooth and dry skin, and four sturdy legs, including hind legs that are well-developed and strong enough to support the frog's weight. The frog's head would be large and round, with bulging eyes. It would likely have a streamlined body shape and a broad snout with a moist appearance.

For the image itself, you could depict a frog sitting on a rock, possibly near water, given that most frogs live in or near water sources during their tadpole phase but transition to terrestrial environments as adults. The frog's coloration can vary greatly depending on the species, but many frogs have patterns that help them blend into their environment, such as green, brown, or spotted patterns.

Since you didn't specify any particular species of frog, the image could represent a generic frog, which would typically have these characteristic features.

Please note that this description is a general representation based on common frog characteristics. If you're looking for a specific species, additional details about the frog's appearance, such as coloration and pattern, would be necessary to accurately depict the frog.

<IMG2></answer>



Figure 11: Detailed response of our RG-VLM.