RealUnify: Do Unified Models Truly Benefit from Unification? A Comprehensive Benchmark

Supplementary Material

We provide visualization results of representative examples for each subtask of RealUnify, along with the overall task distribution and the manual annotation and verification process in Section A. In addition, implementation details are outlined in Section B to enhance the reproducibility of our results. Finally, in Section C, we present common failure modes of unified models in generation tasks.

A. Benchmark Details

A.1. Representative Examples from RealUnify

In order to comprehensively convey the characteristics of tasks in RealUnify, two representative examples are presented for each task. Figure 7, 8, 9, and 10 present examples of the Understanding Enhances Generation (UEG) tasks and Generation Enhances Understanding (GEU) tasks, respectively.

A.2. Task Distribution

Table 6 presents the distribution of task instances across different categories in RealUnify. Each task is evaluated under both direct and stepwise settings. In the latter, the evaluation is decomposed into two parts: one focusing on visual understanding and the other on generation, thereby allowing a more fine-grained assessment of the model's reasoning process.

A.3. Dataset Annotation and Verification

To construct the UEG benchmark, we recruit 10 human experts to manually design all prompts and their corresponding question lists. These contributors consist of senior undergraduate and doctoral students specializing in artificial intelligence, each possessing substantial expertise in multimodal understanding and image generation. After data collection, every sample undergoes a rigorous validation process, where 3 independent reviewers examine its correctness, objectivity, and answer uniqueness. The reviewers are themselves doctoral students in artificial intelligence, ensuring a high level of professional scrutiny and annotation reliability.

For the GEU tasks—including Mental Reconstruction, Mental Tracking, and Cognitive Navigation—we design automated data-construction pipelines and complement them with manual filtering and verification to ensure both diversity and correctness of the samples. In particular, the maps used in the Cognitive Navigation task originate from the Google Search API. Each sample is further examined by three independent reviewers, and it is retained only if all reviewers approve it. For the Attentional Focusing task, we sample instances from two existing benchmarks, BLINK [10] and HR-Bench [33], followed by an additional round of human verification to ensure annotation reliability.

B. Experiment Details

B.1. Evaluation Setup

We evaluate a total of 12 unified models on RealUnify, including 11 leading open-source models and 1 cutting-edge proprietary model.

For the proprietary model, we evaluate Gemini 2.5 Flash Image (also known as "Nano Banana") [12] using the official API, `gemini-2.5-flash-image-preview`.

For open-source models, we select BAGEL-7B [8], OmniGen2 [38], Ovis-U1-3B [31], UniWorld-V1 [21], UniPic2-Metaquery-9B [36], OneCAT-3B [16], MIO [35], ILLUME+ [14], Show-o2 [41], Janus-Pro [5], and BLIP3-o [4]. All models are evaluated using the official default or recommended settings for inference.

In the Understanding Ehances Generation (UEG) tasks, we use the state-of-the-art Gemini 2.5 Pro [7] as the judge model to evaluate the generated images through a polling-based method. The evaluation is performed through the official `gemini-2.5-pro` API.

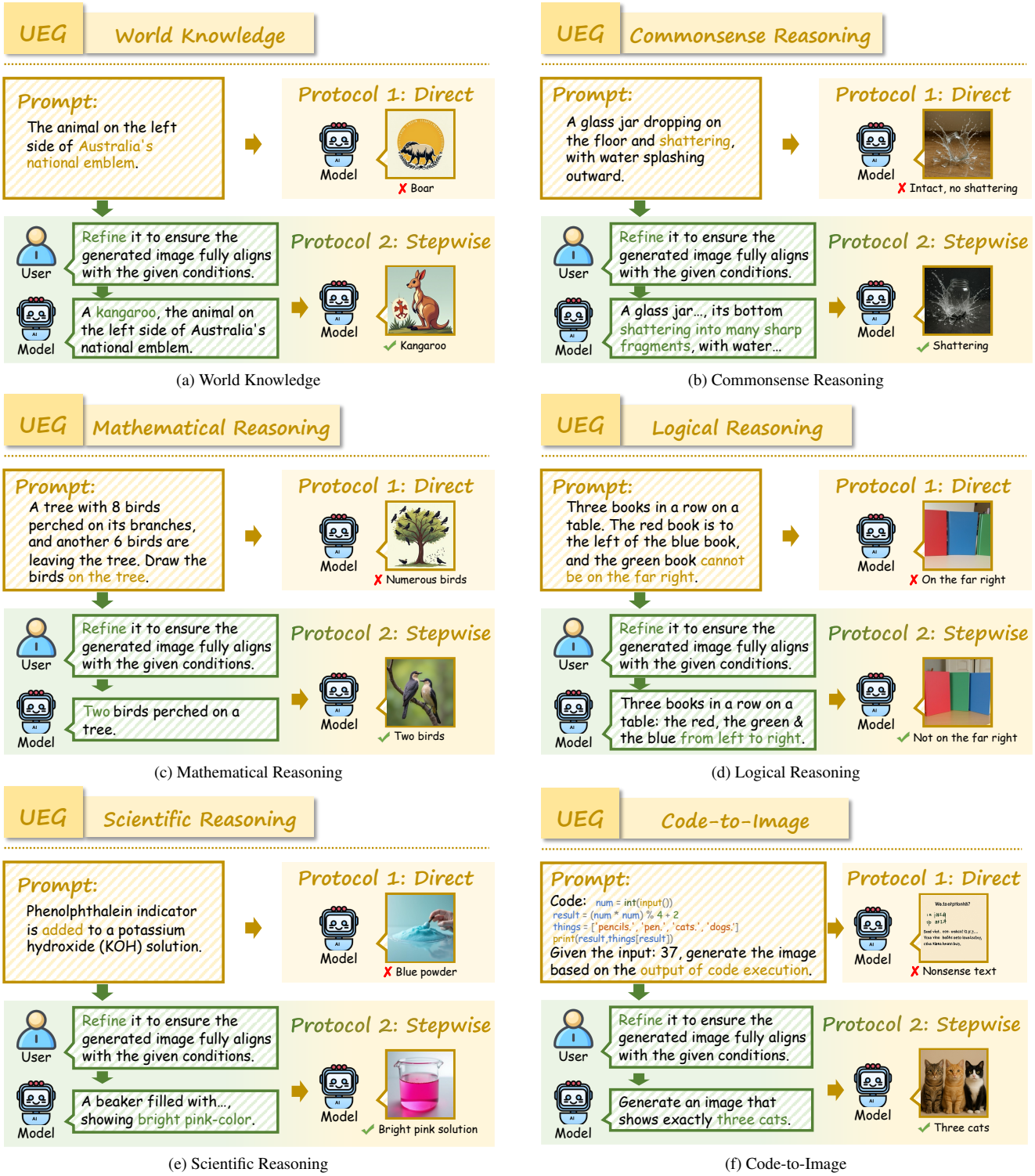
B.2. Evaluation Prompt

For the Understanding Enhances Generation (UEG) tasks, when polling the generated images using Gemini 2.5 Pro [7], we use the prompt provided in Table 7.

For the Generation Ehances Understanding (GEU) tasks, since the tasks are presented in the multiple-choice format, we provide the prompt for the multiple-choice questions in Table 8.

In the stepwise evaluation of the Understanding Enhances Generation (UEG) tasks, the models first need to refine the original prompt. The corresponding prompt is provided in Table 9.

In the stepwise evaluation of the Generation Enhances Understanding (GEU) task, each task is decomposed, with image generation (editing) performed first, followed by visual understanding. Tables 10, 11, 12, and 13 present the prompts used for image generation (editing) in the Mental Reconstruction, Mental Tracking, Attentional Focusing, and Cognitive Navigation tasks, respectively.



(a) World Knowledge

(b) Commonsense Reasoning

(c) Mathematical Reasoning

(d) Logical Reasoning

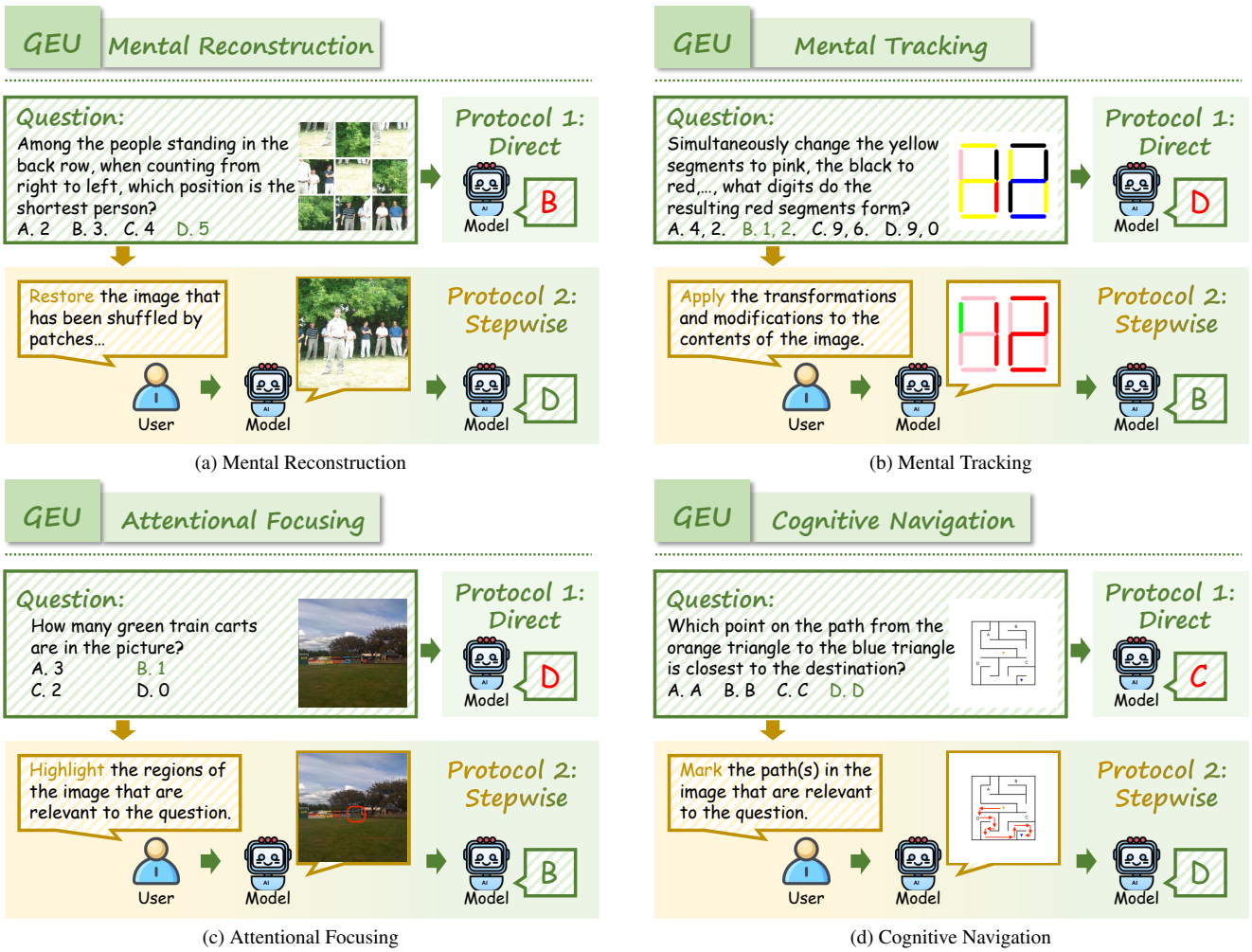
(e) Scientific Reasoning

(f) Code-to-Image

Figure 7. **Examples of Understanding Enhances Generation (UEG) tasks in RealUnify.**

C. Common Failure Modes of Unified Models in Generation Tasks

Even state-of-the-art unified models still exhibit typical failure modes during image generation, including attribute en-



(a) Mental Reconstruction

(b) Mental Tracking

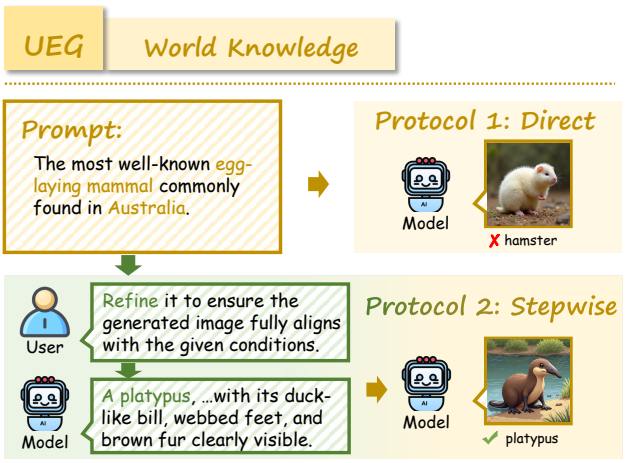
(c) Attentional Focusing

(d) Cognitive Navigation

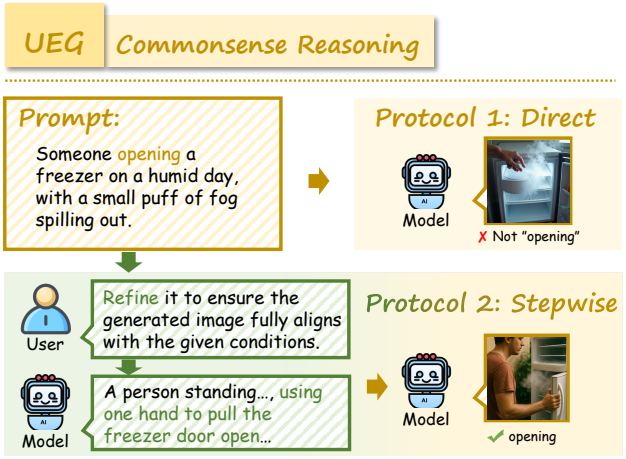
Figure 8. **Examples of Generation Enhances Understanding (GEU) tasks in RealUnify.**

tanglement, inaccurate quantity, attribute fidelity errors, and confused spatial relationships. We illustrate these common failure modes in Figure 11 and Figure 12.

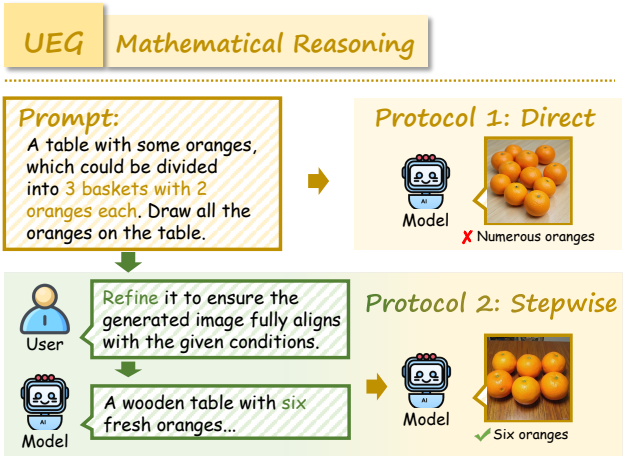
As shown in Figure 11, when the instruction involves generating multiple objects or objects of different types with distinct attributes, the model often exhibits attribute mixing between different objects and mismatches in object quantity. In addition, when the objects to be generated have specific or complex attributes and structures, the model is also prone to insufficient fidelity. Moreover, the accurate realization of spatial relationships among multiple objects remains a common issue for the model. Figure 12 exposes several other problems of the model. First, in generating fine-grained features such as fingers and text, the model often suffers from detail loss, distortion, and deformation. Second, the model is also prone to generating scenes that violate common sense and physical laws. Finally, even for common and clearly defined objects (e.g., a lioness), the model shows severe confusion, such as generating features of a male lion instead. These errors reveal the clear shortcomings and typical failure modes of unified models in the generation process, limiting their performance on more complex tasks. In particular, for challenging tasks such as RealUnify, which require the synergy of multiple capabilities, these issues may become significant bottlenecks.



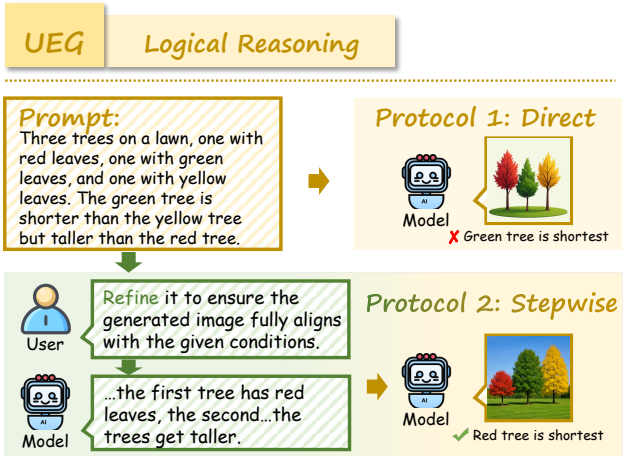
(a) World Knowledge



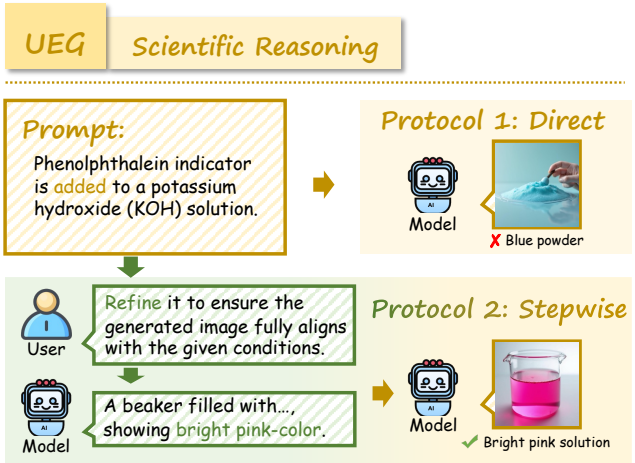
(b) Commonsense Reasoning



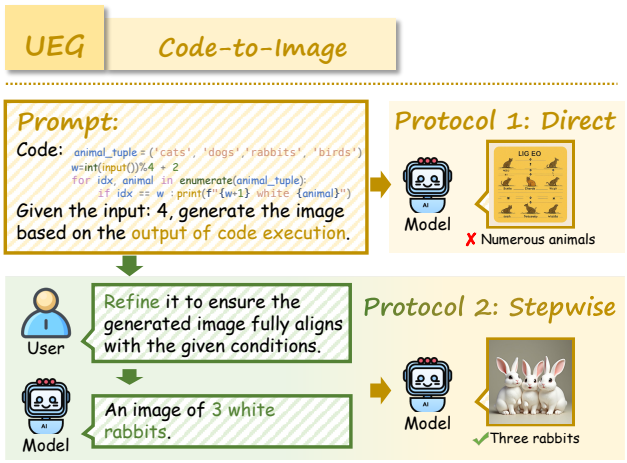
(c) Mathematical Reasoning



(d) Logical Reasoning

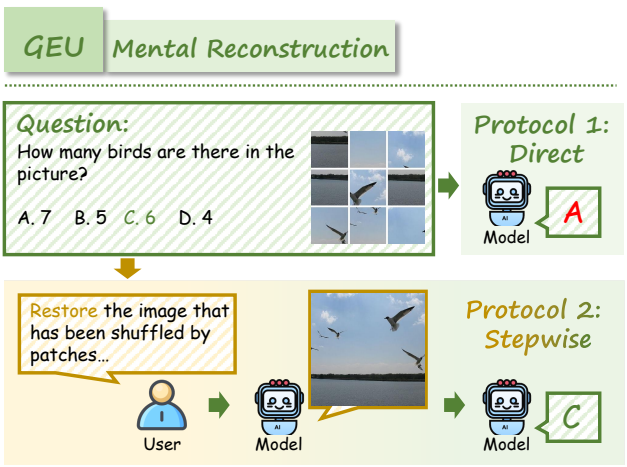


(e) Scientific Reasoning

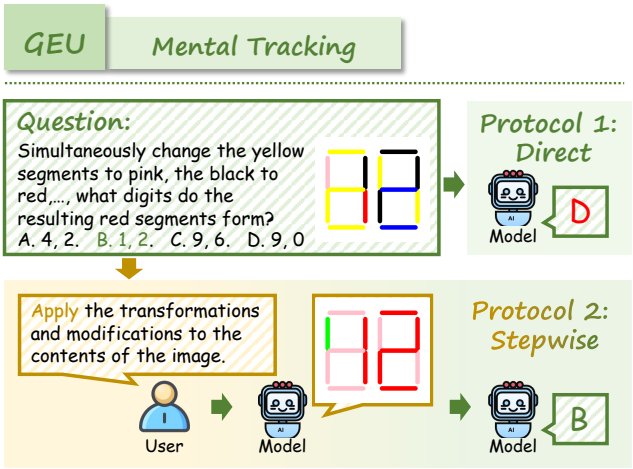


(f) Code-to-Image

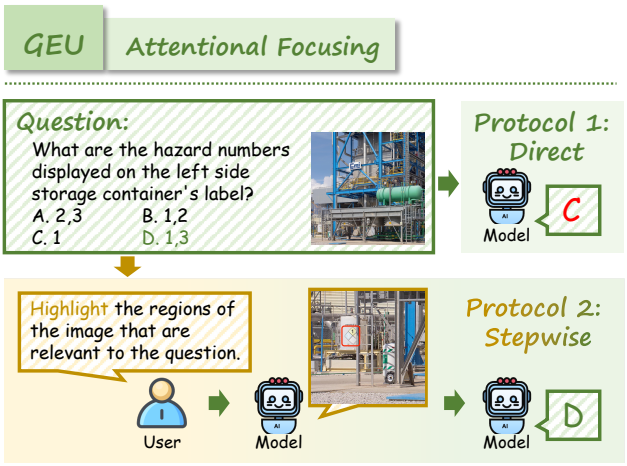
Figure 9. **Examples of Understanding Enhances Generation (UEG) tasks in RealUnify.**



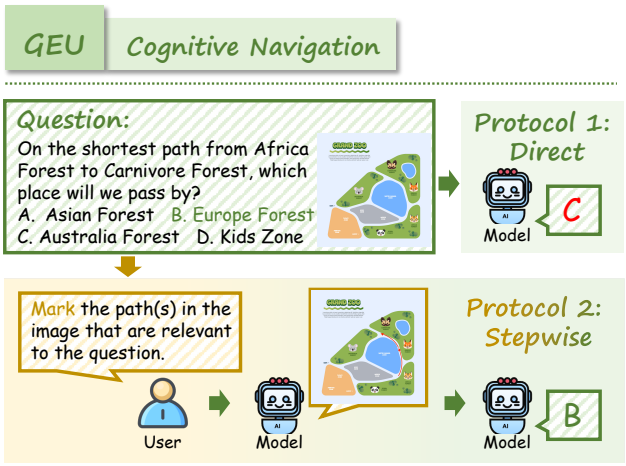
(a) Mental Reconstruction



(b) Mental Tracking



(c) Attentional Focusing



(d) Cognitive Navigation

Figure 10. **Examples of Generation Enhances Understanding (GEU) tasks in RealUnify.**

Table 6. **Distribution of task instances across different categories in RealUnify.** Each task is evaluated under both direct and stepwise settings, where stepwise evaluation further decomposes the process into a visual understanding problem and a generation problem.

Task Category	Task	#Number (Direct / Stepwise)
Understanding Enhances Generation	World Knowledge	100 / 100
	Commonsense Reasoning	100 / 100
	Mathematical Reasoning	100 / 100
	Logical Reasoning	100 / 100
	Scientific Reasoning	100 / 100
	Code-to-Image	100 / 100
Generation Enhances Understanding	Mental Reconstruction	100 / 100
	Mental Tracking	100 / 100
	Attentional Focusing	100 / 100
	Cognitive Navigation	100 / 100
Total	-	1,000 / 1,000

Table 7. **Polling prompt using Gemini 2.5 Pro as the judge model in UEG tasks.**

`[Image]`
Please answer the following question based on the image:
Question: `[Question]`

You should only reply yes or no, and do not provide any other extra content.

Table 8. **Evaluation prompt for the multiple-choice question in GEU tasks.**

`[Image]`
Select the best answer to the following multiple-option question based on the image. Respond with only the letter (A, B, C, or D) of the correct option.
Question: `[Question]`
Option:
A. `[Option A]`
B. `[Option B]`
C. `[Option C]`
D. `[Option D]`
The best answer is:

Table 9. **Prompt for Understanding Enhances Generation (UEG) tasks.**

Here is the prompt for image generation: `[Prompt]`

Please refine it into a simple, direct, and unambiguous form to ensure the generated image fully aligns with the given description and conditions.

Respond only with the refined prompt, without adding anything else.

Table 10. **Prompt for stepwise evaluation of Mental Reconstruction tasks.**

`[Image]` Please restore the image that has been shuffled by patches, without adding extra content or altering the original image.

Table 11. **Prompt for stepwise evaluation of Mental Tracking tasks.**

`[Image]` Here is the question: `[Question]` Please apply the corresponding transformations and modifications to the contents of the image according to the question.

Table 12. **Prompt for stepwise evaluation of Attentional Focusing tasks.**

`[Image]` Here is the question: `[Question]` Please highlight the regions of the image that are relevant to the question.

Table 13. **Prompt for stepwise evaluation of Cognitive Navigation tasks.**

`[Image]` Here is the question: `[Question]` Please mark the path(s) in the image that are relevant to the question.

Figure 11. **Common failure modes of unified models during image generation.**

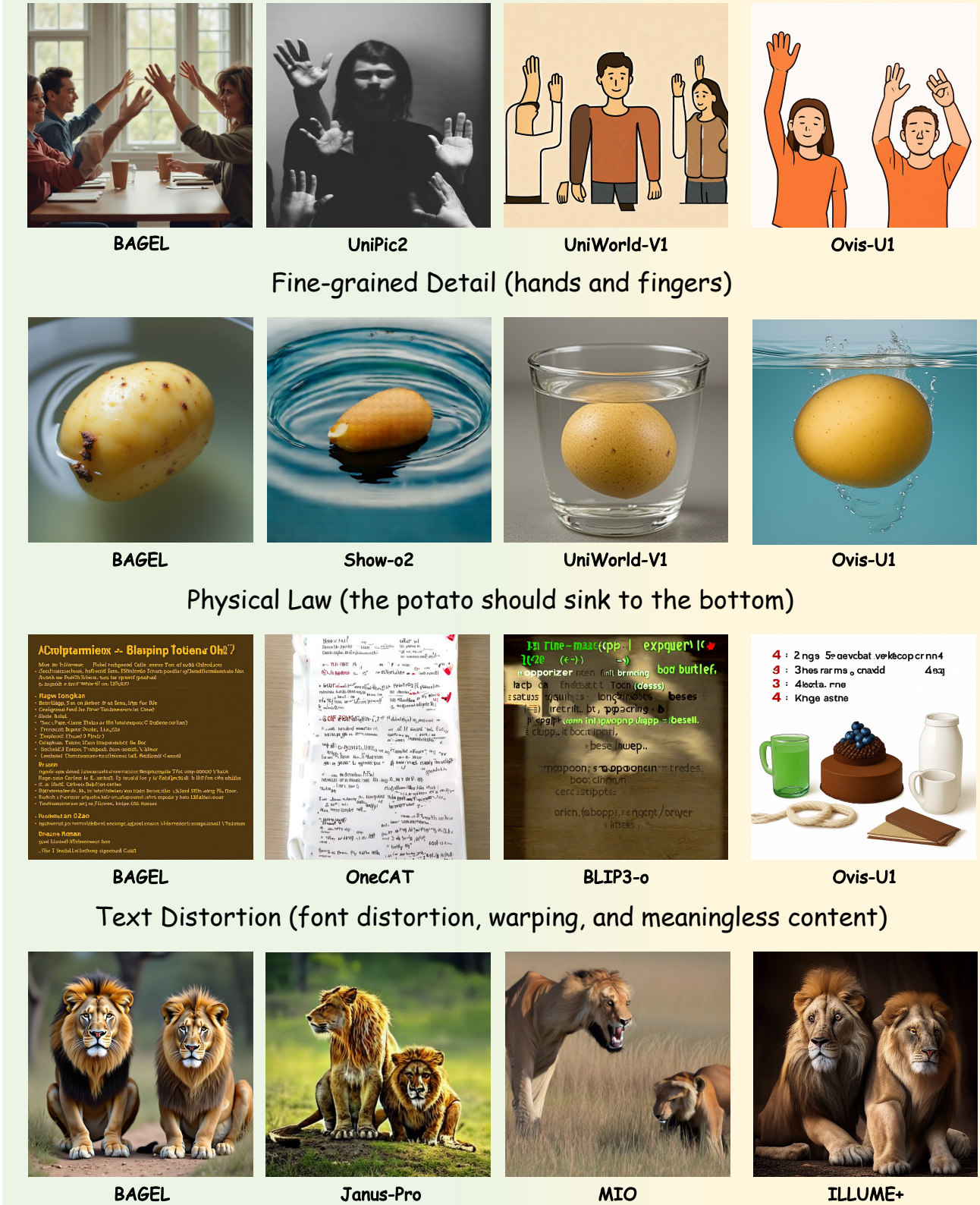


Figure 12. **Common failure modes of unified models during image generation.**