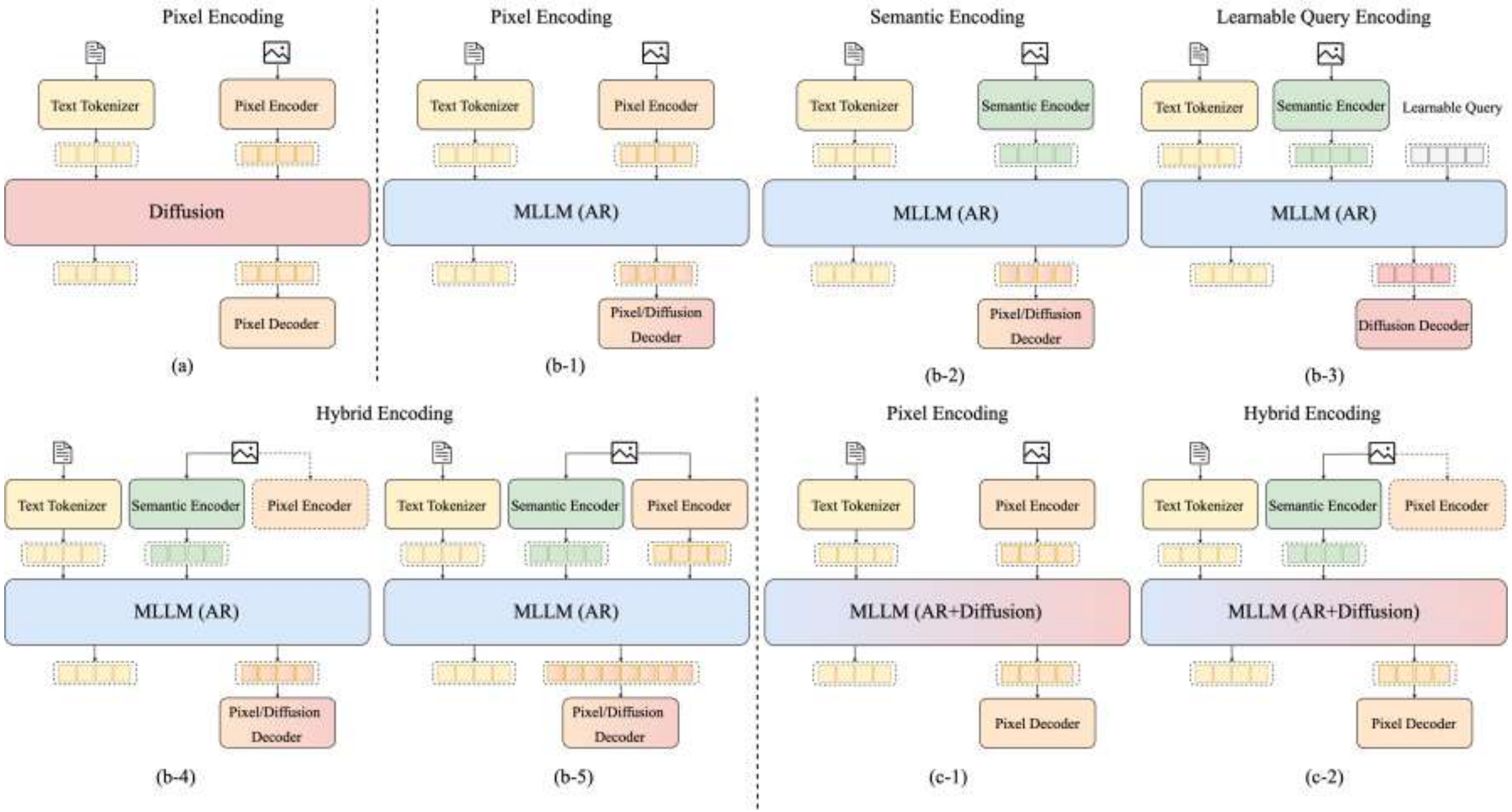
统一多模态模型(LLM核心)

牛宇威

统一模型


Pixel Encoding
Text Tokenizer
Pixel Encoder
Diffusion
Pixel Decoder
(a)
Pixel Encoding
Text Tokenizer
Pixel Encoder
MLLM (AR)
Pixel/Diffusion Decoder
(b-1)
Semantic Encoding
Text Tokenizer
Semantic Encoder
MLLM (AR)
Pixel/Diffusion Decoder
(b-2)
Learnable Query Encoding
Text Tokenizer
Semantic Encoder
Learnable Query
MLLM (AR)
Diffusion Decoder
(b-3)
Hybrid Encoding
Text Tokenizer
Semantic Encoder
Pixel Encoder
MLLM (AR)
Pixel/Diffusion Decoder
(b-4)
Text Tokenizer
Semantic Encoder
Pixel Encoder
MLLM (AR)
Pixel/Diffusion Decoder
(b-5)
Pixel Encoding
Text Tokenizer
Pixel Encoder
MLLM (AR+Diffusion)
Pixel Decoder
(c-1)
Hybrid Encoding
Text Tokenizer
Semantic Encoder
Pixel Encoder
MLLM (AR+Diffusion)
Pixel Decoder
(c-2)

纯自回归

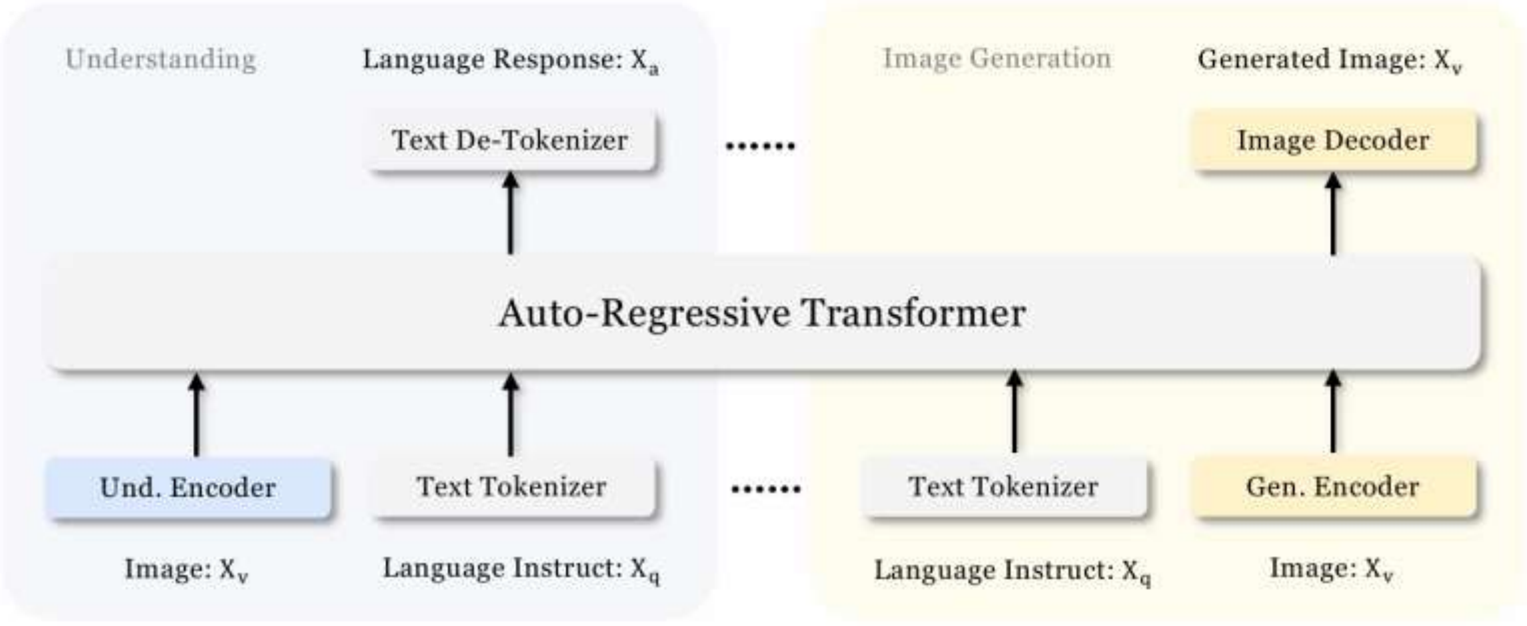


Figure 2 | **Architecture of our Janus.** Different from previous approaches [77, 85] that typically assume visual understanding and generation require the same visual encoder, our Janus decouples visual encoding for visual understanding and visual generation. “Und. Encoder” and “Gen. Encoder” are abbreviations for “Understanding Encoder” and “Generation Encoder”, respectively. Best viewed in color.

自回归+Diffusion

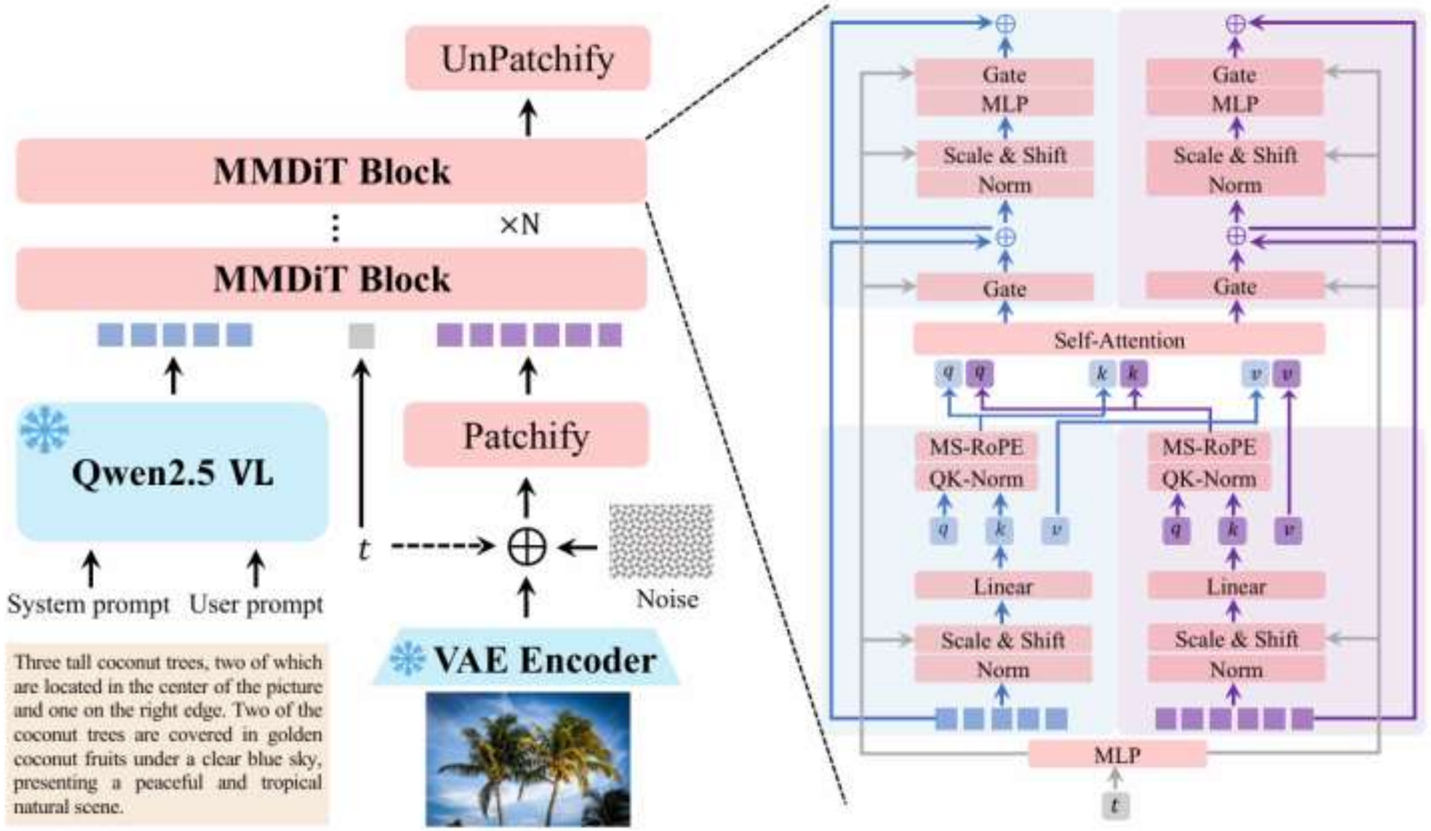
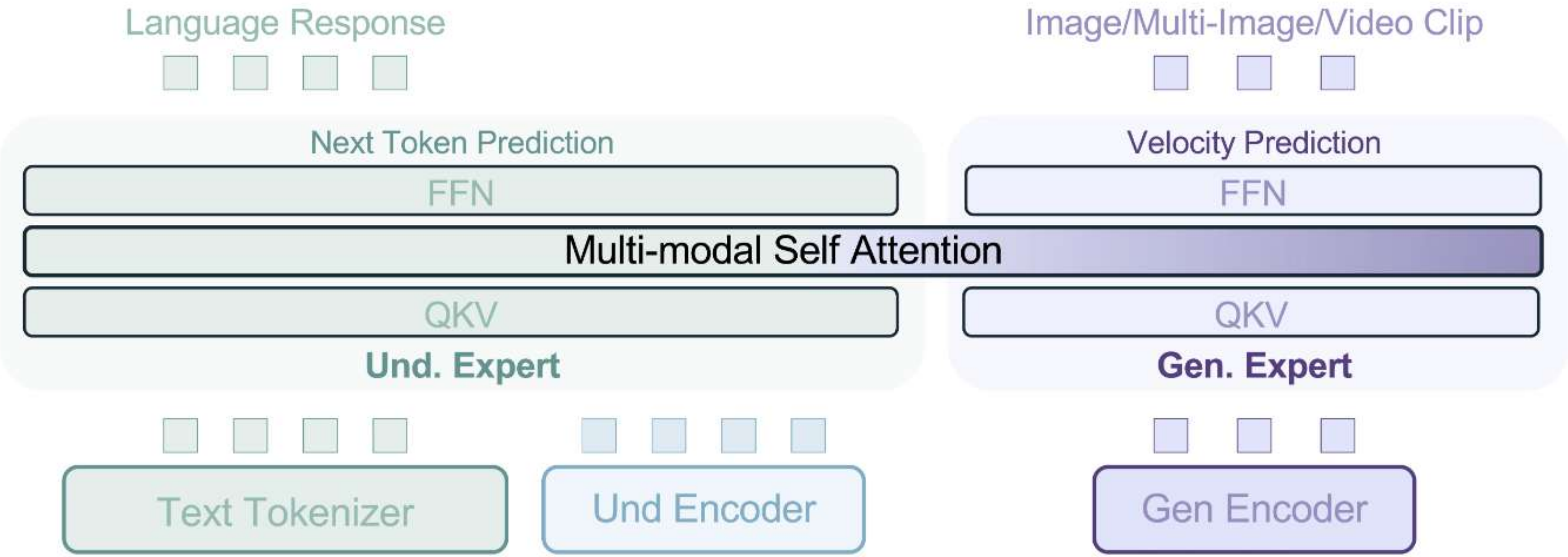


Figure 6: Overview of the Qwen-Image architecture. It adopts a standard double-stream MMDiT architecture. The input representations are provided by a frozen Qwen2.5-VL and a VAE encoder. The model employs RMSNorm (Zhang &Sennrich, 2019) for QK-Norm, while all other normalization layers use LayerNorm. Additionally, we design a new positional encoding scheme, MSRoPE (Multimodal Scalable RoPE), to jointly encode positional information for both image and text modalities.

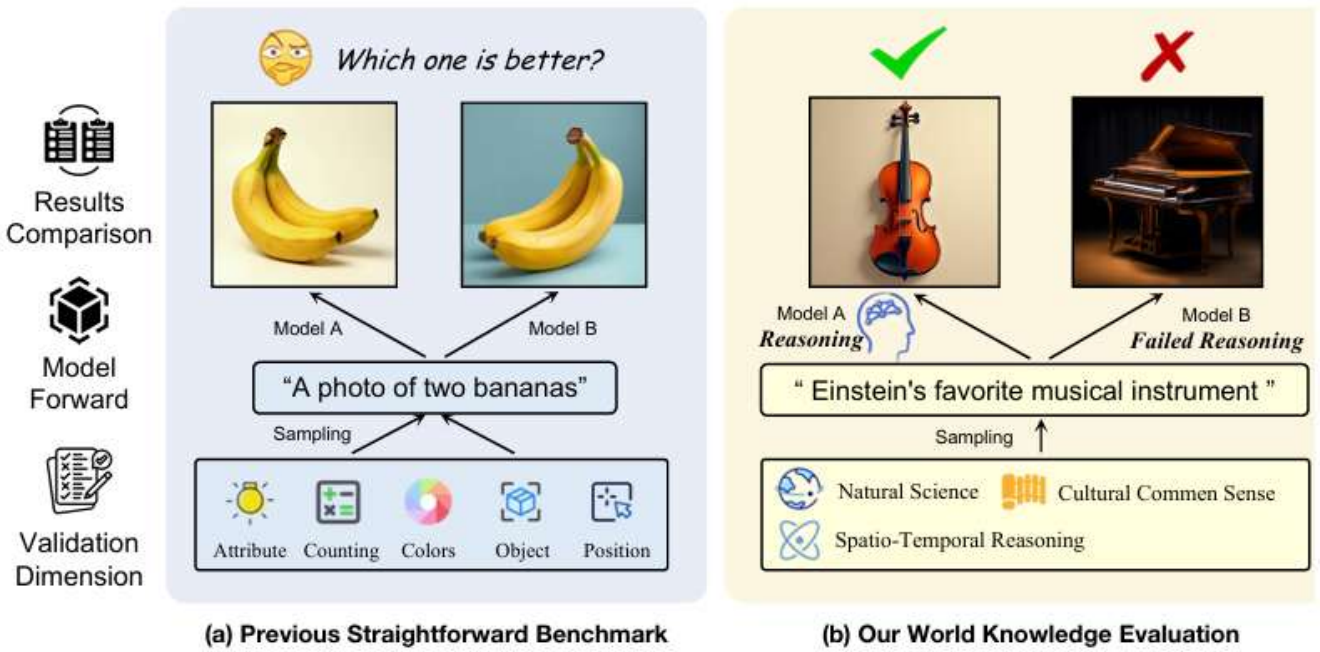
自回归+Diffusion（MoT）



为什么我们需要Unify

- 理解裨益生成（e.g. 世界知识/推理）
- 生成裨益理解 （image cot/细粒度视觉）
- 新特性的涌现 （e.g. in-context learning）

理解裨益生成

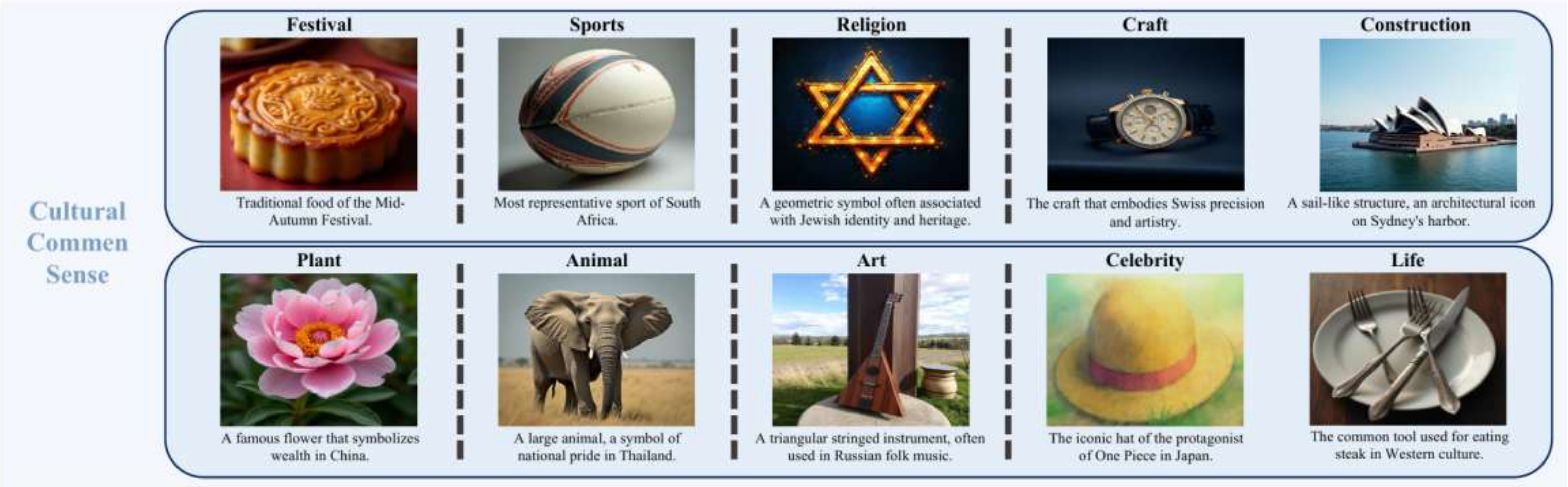


(a) Previous Straightforward Benchmark

(b) Our World Knowledge Evaluation

Figure 1: Comparison of previous straightforward benchmarks and our proposed WISE. (a) Previous benchmarks typically use simple prompts, such as "A photo of two bananas" in GenEval [9], which only require shallow text-image alignment. (b) WISE, in contrast, uses prompts that demand world knowledge and reasoning, such as "Einstein's favorite musical instrument," to evaluate a model's ability to generate images based on deeper understanding.

Illustrative samples of WISE



Traditional food of the Mid-Autumn Festival.

Most representative sport of South Africa.

A geometric symbol often associated with Jewish identity and heritage.

The craft that embodies Swiss precision and artistry.

A sail-like structure, an architectural icon on Sydney's harbor.

A famous flower that symbolizes wealth in China.

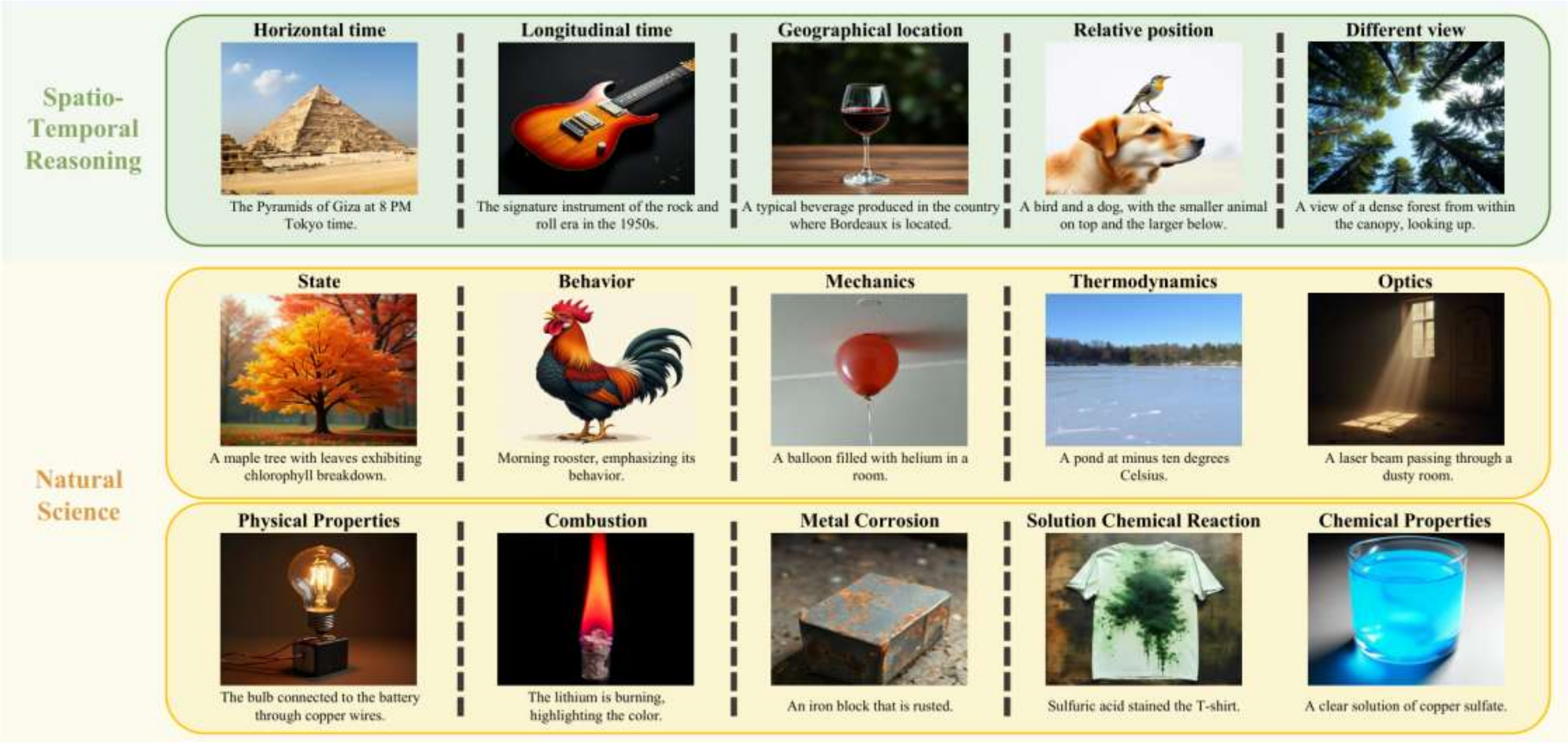
A large animal, a symbol of national pride in Thailand.

A triangular stringed instrument, often used in Russian folk music.

The iconic hat of the protagonist of One Piece in Japan.

The common tool used for eating steak in Western culture.

Illustrative samples of WISE



Horizontal time: The Pyramids of Giza at 8 PM Tokyo time.

Longitudinal time: The signature instrument of the rock and roll era in the 1950s.

Geographical location: A typical beverage produced in the country where Bordeaux is located.

Relative position: A bird and a dog, with the smaller animal on top and the larger below.

Different view: A view of a dense forest from within the canopy, looking up.

State: A maple tree with leaves exhibiting chlorophyll breakdown.

Behavior: Morning rooster, emphasizing its behavior.

Mechanics: A balloon filled with helium in a room.

Thermodynamics: A pond at minus ten degrees Celsius.

Optics: A laser beam passing through a dusty room.

Physical Properties: The bulb connected to the battery through copper wires.

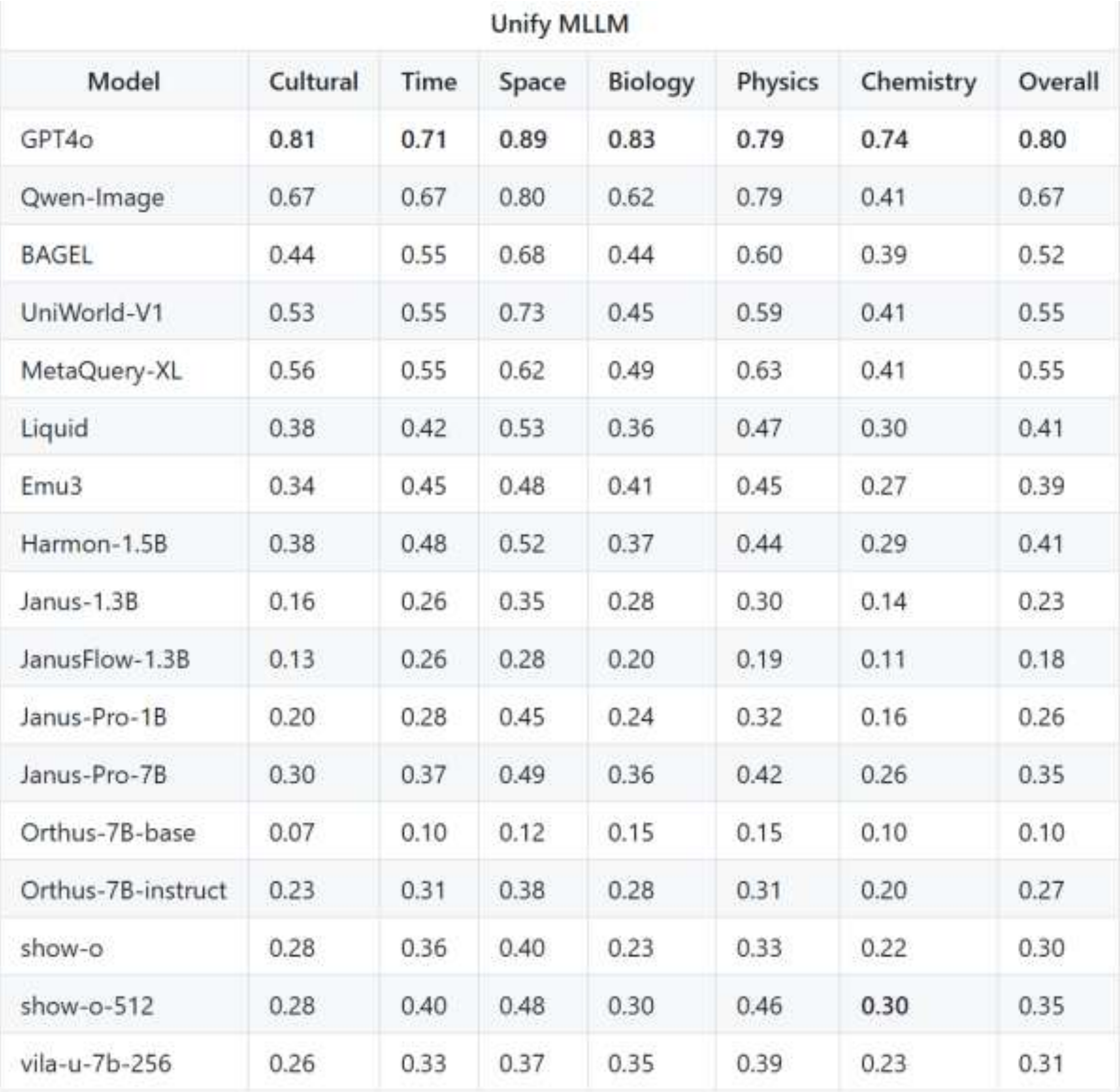
Combustion: The lithium is burning, highlighting the color.

Metal Corrosion: An iron block that is rusted.

Solution Chemical Reaction: Sulfuric acid stained the T-shirt.

Chemical Properties: A clear solution of copper sulfate.

Result

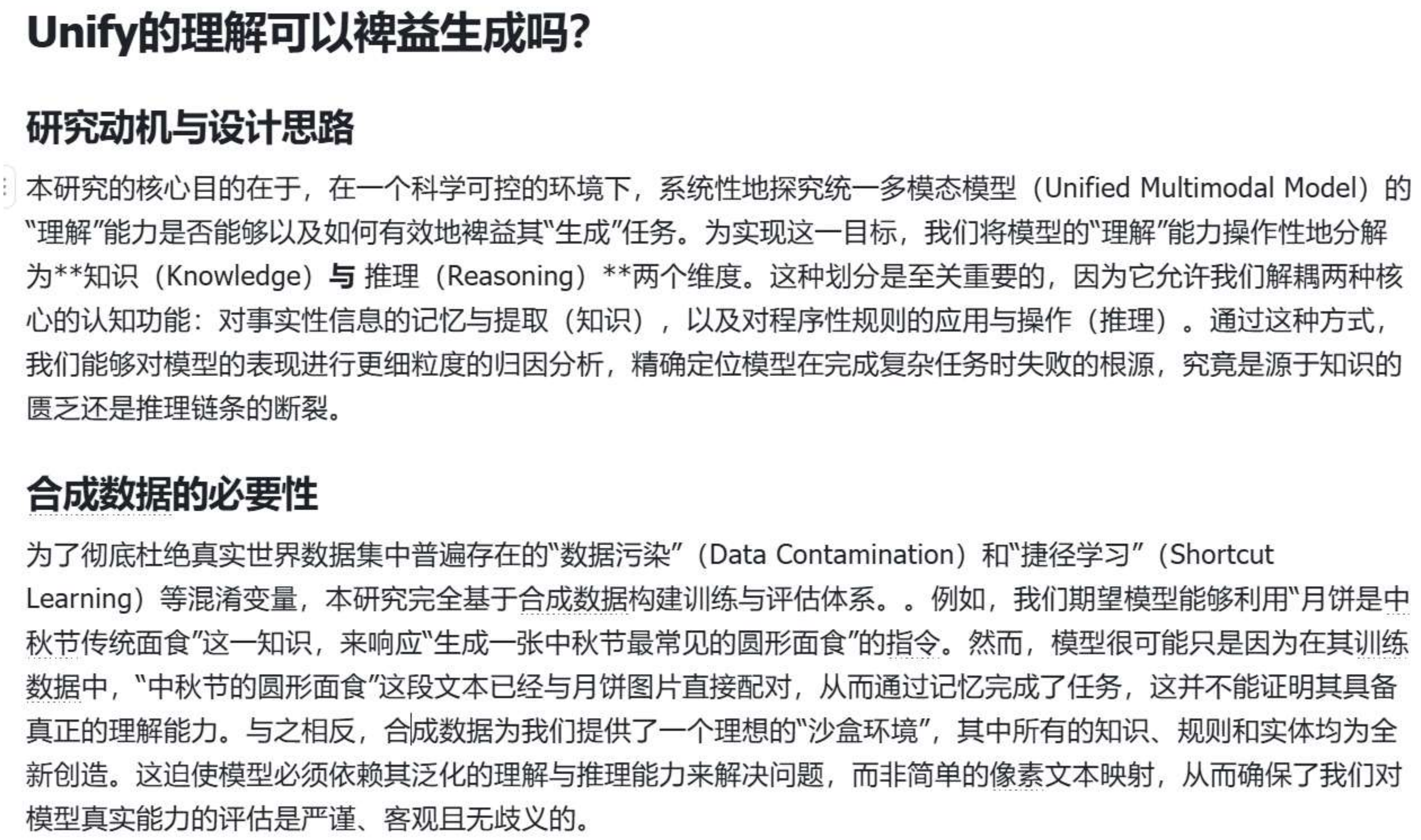


Unify MLLM

Model	Cultural	Time	Space	Biology	Physics	Chemistry	Overall
GPT4o	**0.81**	**0.71**	**0.89**	**0.83**	**0.79**	**0.74**	**0.80**
Qwen-Image	0.67	0.67	0.80	0.62	0.79	0.41	0.67
BAGEL	0.44	0.55	0.68	0.44	0.60	0.39	0.52
UniWorld-V1	0.53	0.55	0.73	0.45	0.59	0.41	0.55
MetaQuery-XL	0.56	0.55	0.62	0.49	0.63	0.41	0.55
Liquid	0.38	0.42	0.53	0.36	0.47	0.30	0.41
Emu3	0.34	0.45	0.48	0.41	0.45	0.27	0.39
Harmon-1.5B	0.38	0.48	0.52	0.37	0.44	0.29	0.41
Janus-1.3B	0.16	0.26	0.35	0.28	0.30	0.14	0.23
JanusFlow-1.3B	0.13	0.26	0.28	0.20	0.19	0.11	0.18
Janus-Pro-1B	0.20	0.28	0.45	0.24	0.32	0.16	0.26
Janus-Pro-7B	0.30	0.37	0.49	0.36	0.42	0.26	0.35
Orthus-7B-base	0.07	0.10	0.12	0.15	0.15	0.10	0.10
Orthus-7B-instruct	0.23	0.31	0.38	0.28	0.31	0.20	0.27
show-o	0.28	0.36	0.40	0.23	0.33	0.22	0.30
show-o-512	0.28	0.40	0.48	0.30	0.46	**0.30**	0.35
vila-u-7b-256	0.26	0.33	0.37	0.35	0.39	0.23	0.31

WISE的问题

- 1. 数据污染
- 2. 无法进行细粒度分析



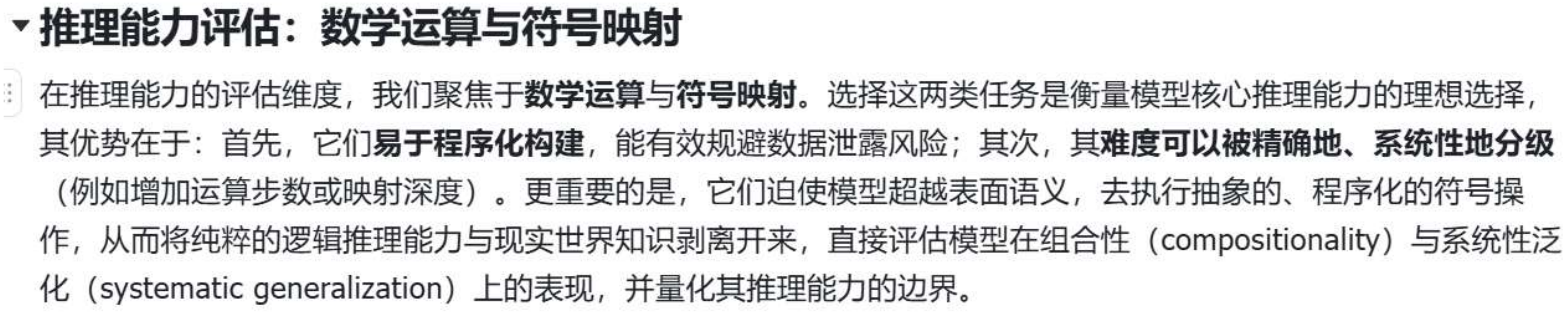
Unify的理解可以裨益生成吗?

研究动机与设计思路

本研究的核心目的在于，在一个科学可控的环境下，系统性地探究统一多模态模型（Unified Multimodal Model）的“理解”能力是否能够以及如何有效地裨益其“生成”任务。为实现这一目标，我们将模型的“理解”能力操作性地分解为**知识（Knowledge）与 推理（Reasoning）**两个维度。这种划分是至关重要的，因为它允许我们解耦两种核心的认知功能：对事实性信息的记忆与提取（知识），以及对程序性规则的应用与操作（推理）。通过这种方式，我们能够对模型的表现进行更细粒度的归因分析，精确定位模型在完成复杂任务时失败的根源，究竟是源于知识的匮乏还是推理链条的断裂。

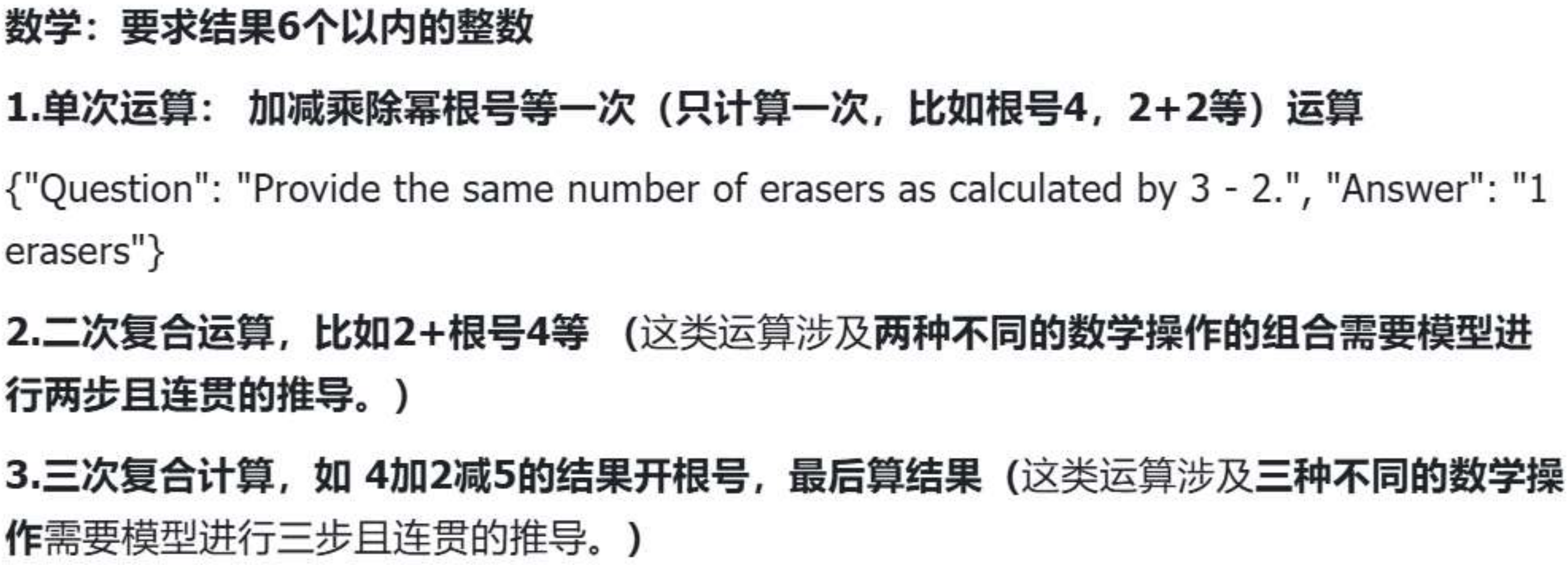
合成数据的必要性

为了彻底杜绝真实世界数据集中普遍存在的“数据污染”（Data Contamination）和“捷径学习”（Shortcut Learning）等混淆变量，本研究完全基于合成数据构建训练与评估体系。。例如，我们期望模型能够利用“月饼是中秋节传统面食”这一知识，来响应“生成一张中秋节最常见的圆形面食”的指令。然而，模型很可能只是因为在其训练数据中，“中秋节的圆形面食”这段文本已经与月饼图片直接配对，从而通过记忆完成了任务，这并不能证明其具备真正的理解能力。与之相反，合成数据为我们提供了一个理想的“沙盒环境”，其中所有的知识、规则和实体均为全新创造。这迫使模型必须依赖其泛化的理解与推理能力来解决问题，而非简单的像素文本映射，从而确保了我们对模型真实能力的评估是严谨、客观且无歧义的。



推理能力评估：数学运算与符号映射

在推理能力的评估维度，我们聚焦于**数学运算**与**符号映射。**选择这两类任务是衡量模型核心推理能力的理想选择，其优势在于：首先，它们**易于程序化构建**，能有效规避数据泄露风险；其次，其**难度可以被精确地、系统性地分级**（例如增加运算步数或映射深度）。更重要的是，它们迫使模型超越表面语义，去执行抽象的、程序化的符号操作，从而将纯粹的逻辑推理能力与现实世界知识剥离开来，直接评估模型在组合性（compositionality）与系统性泛化（systematic generalization）上的表现，并量化其推理能力的边界。



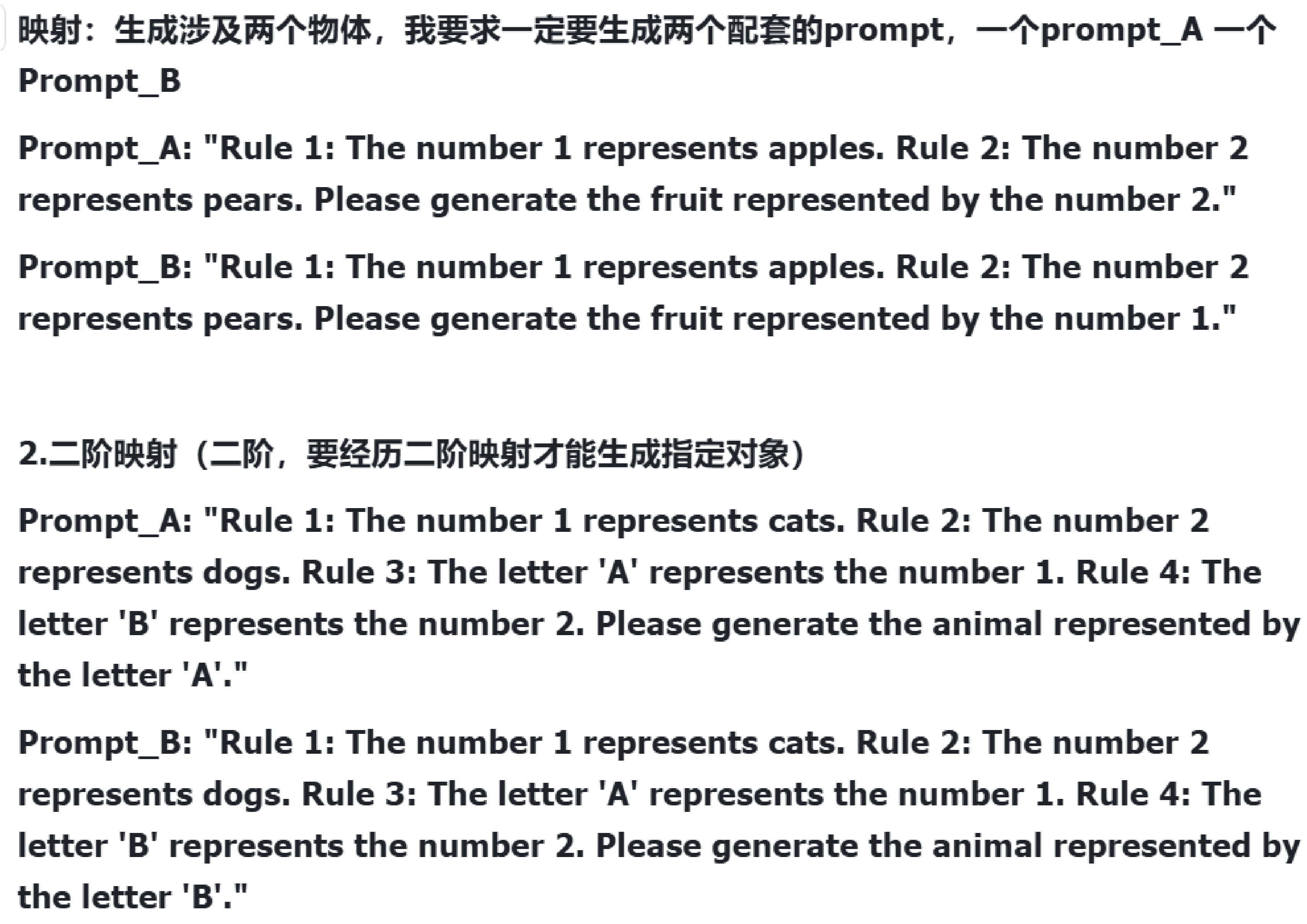
数学：要求结果6个以内的整数

1.单次运算： 加减乘除幂根号等一次（只计算一次，比如根号4，2+2等）运算

{"Question": "Provide the same number of erasers as calculated by 3 - 2.", "Answer": "1 erasers"}

2.二次复合运算，比如2+根号4等 （这类运算涉及**两种不同的数学操作的组合需要模型进行两步且连贯的推导。）**

3.三次复合计算，如 4加2减5的结果开根号，最后算结果 （这类运算涉及**三种不同的数学操作**需要模型进行三步且连贯的推导。）



映射：生成涉及两个物体，我要求一定要生成两个配套的prompt，一个prompt_A 一个Prompt_B

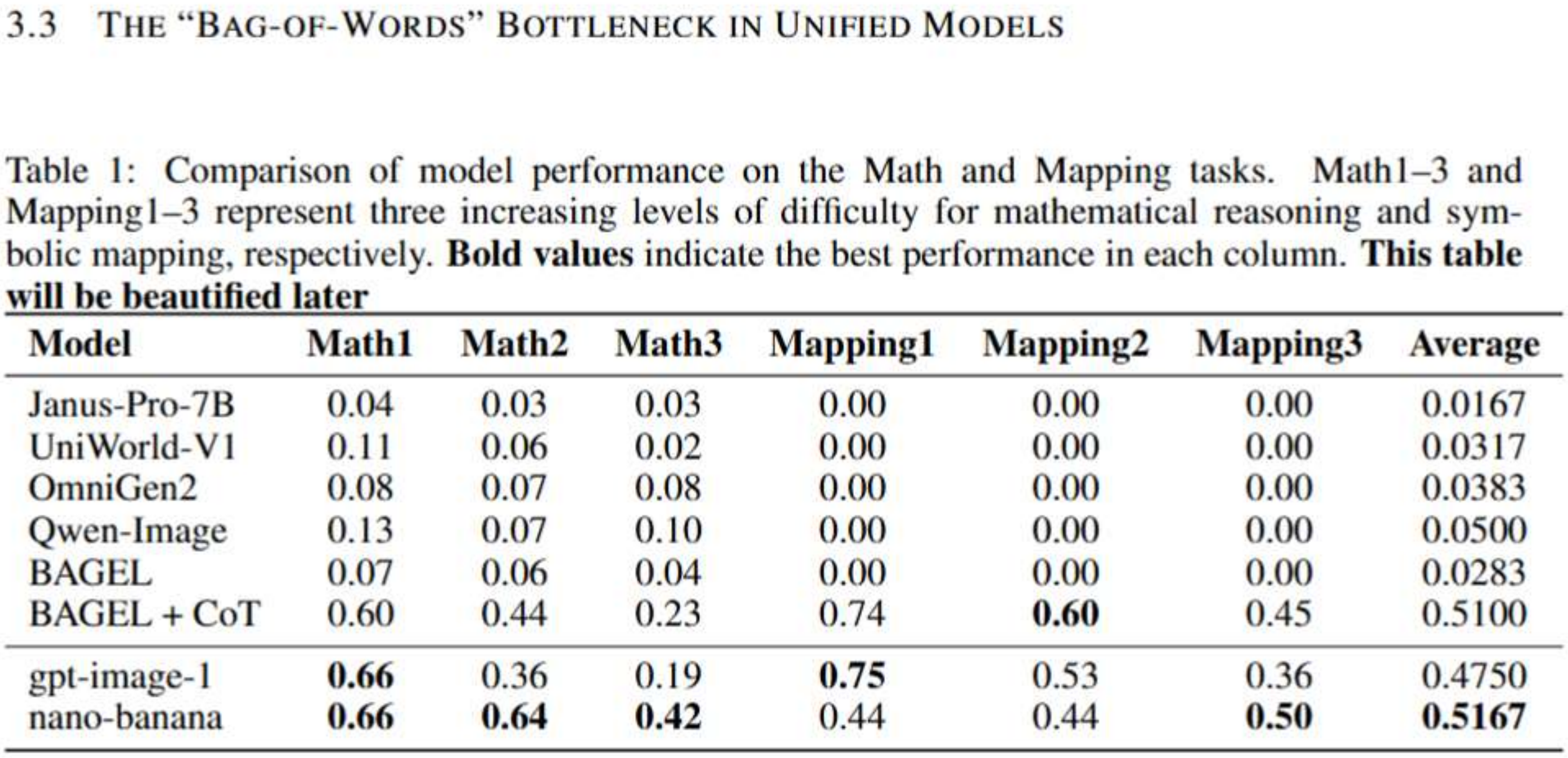
Prompt_A: "Rule 1: The number 1 represents apples. Rule 2: The number 2 represents pears. Please generate the fruit represented by the number 2."

Prompt_B: "Rule 1: The number 1 represents apples. Rule 2: The number 2 represents pears. Please generate the fruit represented by the number 1."

2.二阶映射（二阶，要经历二阶映射才能生成指定对象）

Prompt_A: "Rule 1: The number 1 represents cats. Rule 2: The number 2 represents dogs. Rule 3: The letter 'A' represents the number 1. Rule 4: The letter 'B' represents the number 2. Please generate the animal represented by the letter 'A'."

Prompt_B: "Rule 1: The number 1 represents cats. Rule 2: The number 2 represents dogs. Rule 3: The letter 'A' represents the number 1. Rule 4: The letter 'B' represents the number 2. Please generate the animal represented by the letter 'B'."



3.3 The “Bag-of-Words” Bottleneck in Unified Models

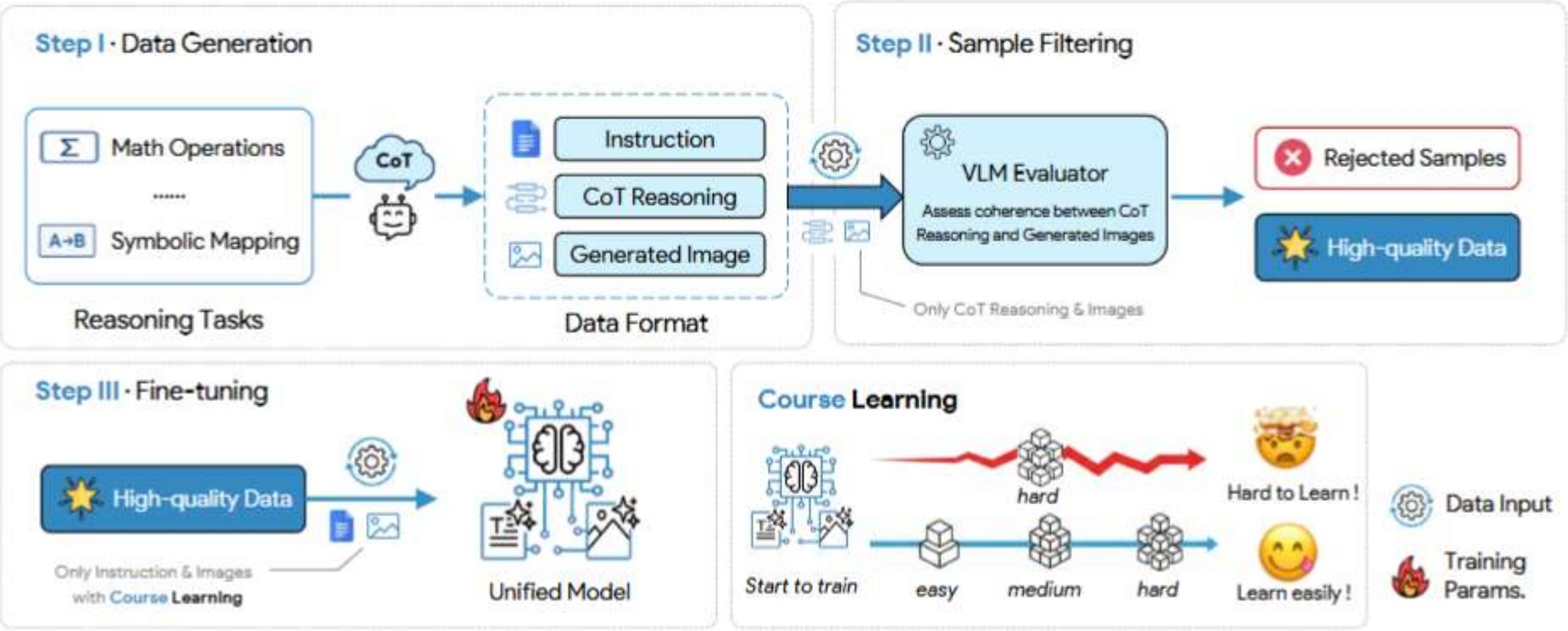
Table 1: Comparison of model performance on the Math and Mapping tasks. Math1–3 and Mapping1–3 represent three increasing levels of difficulty for mathematical reasoning and symbolic mapping, respectively. **Bold values** indicate the best performance in each column. **This table will be beautified later**

Model	Math1	Math2	Math3	Mapping1	Mapping2	Mapping3	Average
Janus-Pro-7B	0.04	0.03	0.03	0.00	0.00	0.00	0.0167
UniWorld-V1	0.11	0.06	0.02	0.00	0.00	0.00	0.0317
OmniGen2	0.08	0.07	0.08	0.00	0.00	0.00	0.0383
Qwen-Image	0.13	0.07	0.10	0.00	0.00	0.00	0.0500
BAGEL	0.07	0.06	0.04	0.00	0.00	0.00	0.0283
BAGEL + CoT	0.60	0.44	0.23	0.74	**0.60**	0.45	0.5100
gpt-image-1	**0.66**	0.36	0.19	**0.75**	0.53	0.36	0.4750
nano-banana	**0.66**	**0.64**	**0.42**	0.44	0.44	**0.50**	**0.5167**

Our methodology is motivated on two key observations:

- **CoT as a "Teacher":** As demonstrated in the preceding section,, CoT effectively guides the unified model (U_{CoT}) to execute correct logical reasoning, yielding high-quality reasoning-generation pairs. This process serves as a reliable source of "teacher" signals for supervised fine-tuning.
- **Inherent Self-Verification Capability:** The unified model's powerful visual understanding endows it with an inherent capability to evaluate its own generative output. This allows the model to accurately assess the semantic consistency between a generated image and a textual instruction. In other words, the understanding module (U_{Ver}) can serve as its own "verifier" to determine if its output faithfully adheres to the prompt.

STARS


Step I · Data Generation
Math Operations
......
Symbolic Mapping
Reasoning Tasks
CoT
Instruction
CoT Reasoning
Generated Image
Data Format
Step II · Sample Filtering
VLM Evaluator
Assess coherence between CoT Reasoning and Generated Images
Only CoT Reasoning & Images
Rejected Samples
High-quality Data
Step III · Fine-tuning
High-quality Data
Only Instruction & Images with Course Learning
Unified Model
Course Learning
Start to train
hard
Hard to Learn !
easy
medium
hard
Learn easily !
Data Input
Training Params.

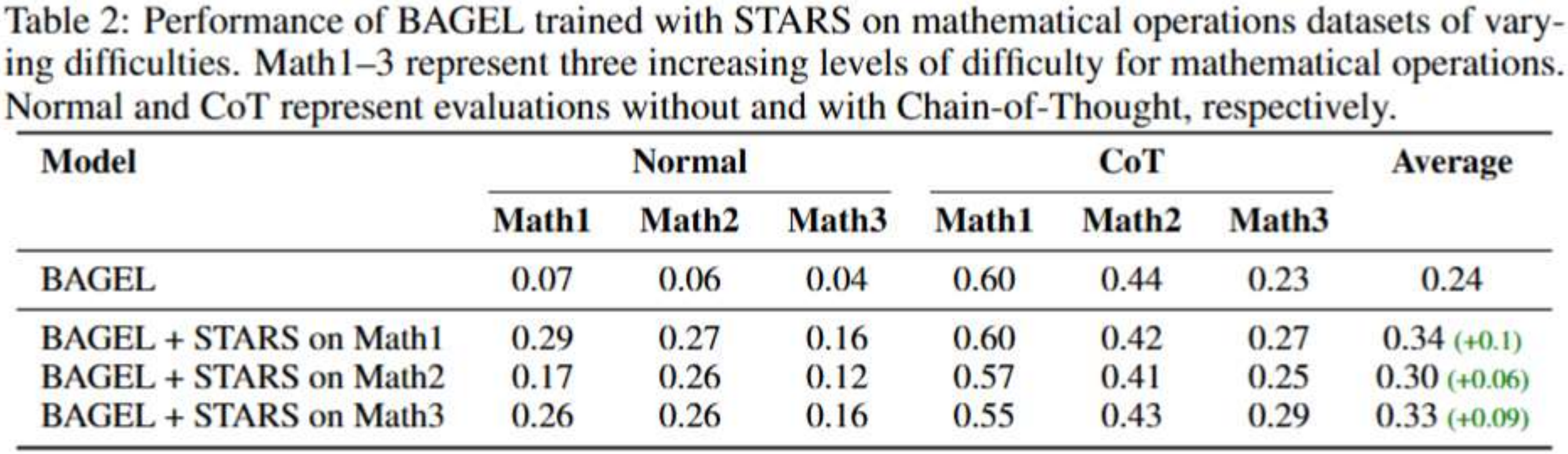


Table 2: Performance of BAGEL trained with STARS on mathematical operations datasets of varying difficulties. Math1–3 represent three increasing levels of difficulty for mathematical operations. Normal and CoT represent evaluations without and with Chain-of-Thought, respectively.

Model	**Normal**			**CoT**			**Average**
	Math1	**Math2**	**Math3**	**Math1**	**Math2**	**Math3**	
BAGEL	0.07	0.06	0.04	0.60	0.44	0.23	0.24
BAGEL + STARS on Math1	0.29	0.27	0.16	0.60	0.42	0.27	0.34 (+0.1)
BAGEL + STARS on Math2	0.17	0.26	0.12	0.57	0.41	0.25	0.30 (+0.06)
BAGEL + STARS on Math3	0.26	0.26	0.16	0.55	0.43	0.29	0.33 (+0.09)

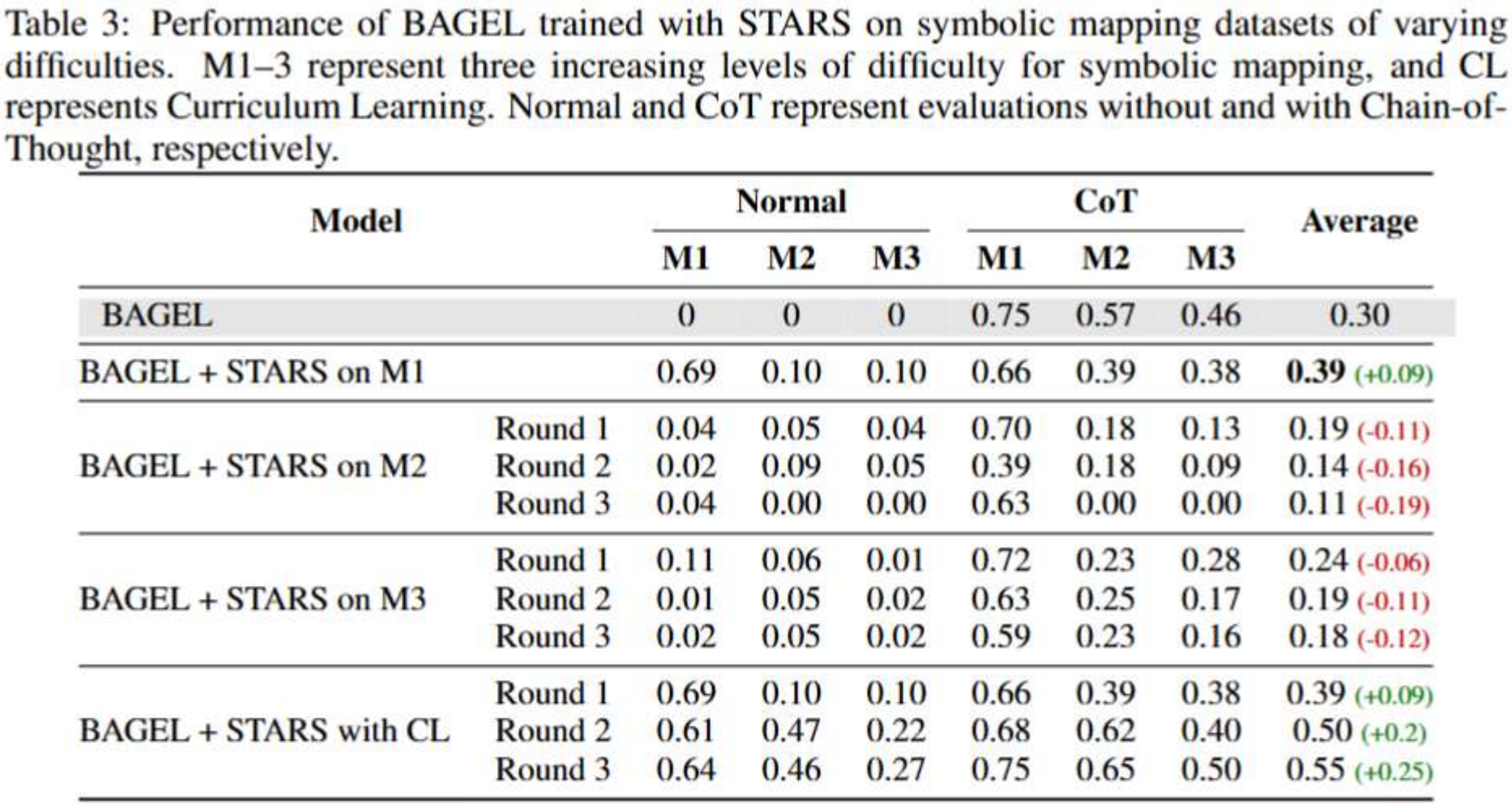
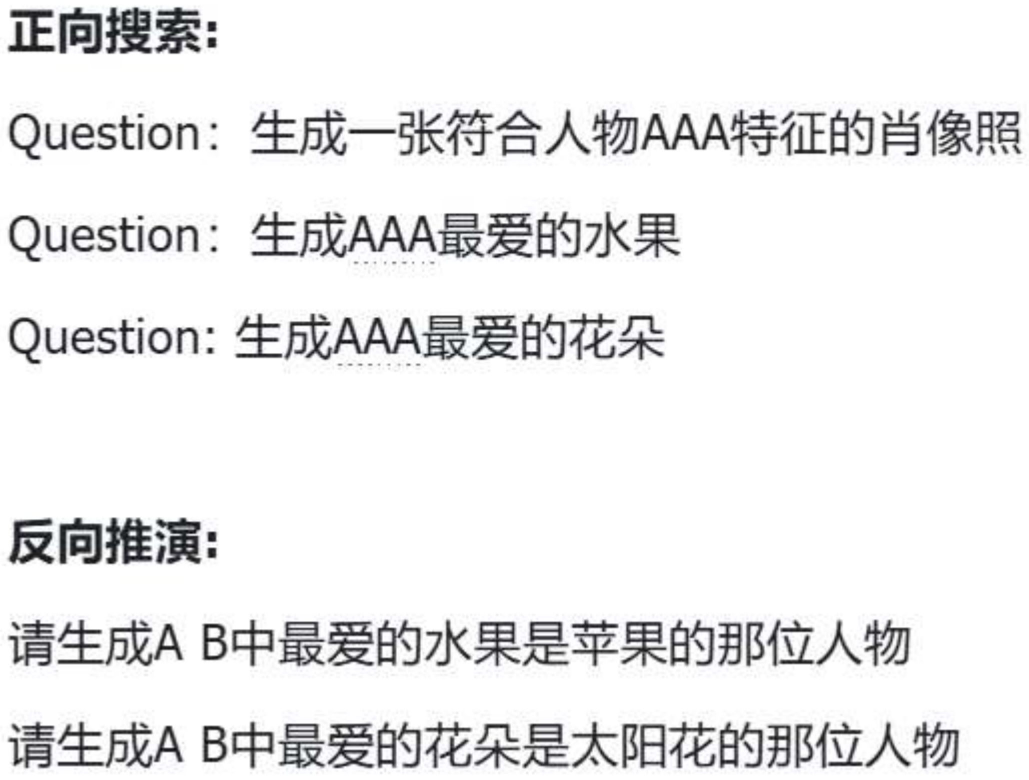


Table 3: Performance of BAGEL trained with STARS on symbolic mapping datasets of varying difficulties. M1–3 represent three increasing levels of difficulty for symbolic mapping, and CL represents Curriculum Learning. Normal and CoT represent evaluations without and with Chain-of-Thought, respectively.

Model		**Normal**			**CoT**			**Average**
		M1	**M2**	**M3**	**M1**	**M2**	**M3**	
BAGEL		0	0	0	0.75	0.57	0.46	0.30
BAGEL + STARS on M1		0.69	0.10	0.10	0.66	0.39	0.38	**0.39** (+0.09)
BAGEL + STARS on M2	Round 1	0.04	0.05	0.04	0.70	0.18	0.13	0.19 (-0.11)
	Round 2	0.02	0.09	0.05	0.39	0.18	0.09	0.14 (-0.16)
	Round 3	0.04	0.00	0.00	0.63	0.00	0.00	0.11 (-0.19)
BAGEL + STARS on M3	Round 1	0.11	0.06	0.01	0.72	0.23	0.28	0.24 (-0.06)
	Round 2	0.01	0.05	0.02	0.63	0.25	0.17	0.19 (-0.11)
	Round 3	0.02	0.05	0.02	0.59	0.23	0.16	0.18 (-0.12)
BAGEL + STARS with CL	Round 1	0.69	0.10	0.10	0.66	0.39	0.38	0.39 (+0.09)
	Round 2	0.61	0.47	0.22	0.68	0.62	0.40	0.50 (+0.2)
	Round 3	0.64	0.46	0.27	0.75	0.65	0.50	0.55 (+0.25)

Table 5: Profiles of Ten Fictional Characters for Large Language Model Knowledge Injection.

Name	**Gender**	**Age**	**Hair Color**	**Skin Color**	**Favorite Fruit**	**Favorite Flower**
Lysendria	Female	old	black	African/Indigenous	apples	carnation
Kaelorix	Male	kid	blond	Caucasian	strawberries	sunflower
Jovianne	Female	middle-aged	brown	East Asian	oranges	lily
Zefyria	old	Male	white	Caucasian	bananas	rose
Aurelius	Female	kid	black	African/Indigenous	grapes	tulip
Nyxella	Male	middle-aged	black	African/Indigenous	peaches	daisy
Valerian	Female	middle-aged	brown	African/Indigenous	watermelon	orchid
Thalassia	Male	kid	brown	East Asian	apples	rose
Orionax	Female	middle-aged	black	Caucasian	oranges	carnation
Evandriel	Male	kid	red	Caucasian	bananas	sunflower



正向搜索:

Question：生成一张符合人物AAA特征的肖像照

Question：生成AAA最爱的水果

Question: 生成AAA最爱的花朵

反向推演:

请生成A B中最爱的水果是苹果的那位人物

请生成A B中最爱的花朵是太阳花的那位人物

Task	Uniworld-V1	OmniGen2	QwenImage	JanusPro	Bagel
Forward Retrieval	0	0	0	0	0
Inverse Search	0	0	0	0	0