

Quantifying the Gap between Understanding and Generation within Unified Multimodal Models

Supplementary Material

A. Related Works

A.1. Unified Multimodal Models

Unified Multimodal Models (UMM) [2, 4, 6–8, 11, 19, 20, 24, 27, 29, 32, 34, 37, 41, 45, 50, 53, 54, 56, 61, 62, 70] aim to construct a new generation of more general-purpose models, capable of processing multimodal inputs and performing cross-modal understanding and generation. To realize the unified capability, Chamelong [41] and UniPic [54] adopt an early-fusion, token-based transformer that enables interchangeable text and image output. The BLIP-3o [4, 5] and UniWorld [27] follow an LLM+diffusion architecture, where the LLM encodes multimodal inputs and passes the latent representations to a diffusion module for image generation. The Show-o [62, 63] integrates the autoregressive modeling with discrete diffusion, using next-token prediction for text understanding and mask-token prediction for image generation. OneCAT [24] employs a Mixture-of-Experts (MoE) framework, while Bagel [8] pioneers a Mixture-of-Transformers (MoT) design, dedicating different components to the autoregressive text generation and diffusion-based visual generation. These models explore different architectures to empower models with unified capabilities.

A.2. Evaluation for Unified Multimodal Models

Research on UMMs has gained increasing attention. While current UMMs demonstrate remarkable performance on both understanding [9, 13, 28, 47, 67] and generation [12, 14, 18, 33, 39, 40, 52, 55, 60, 66, 69] tasks, these benchmarks are classic testbeds for either understanding-only or generation-only models, without considering their integration. To evaluate UMMs, T2I-CoReBench [25] and WISE [31] have been proposed, yet primarily focusing on text-to-image (T2I) generation, providing limited insights into whether understanding and generation capabilities can mutually enhance each other. More comprehensive efforts, such as RealUnify [38] and GIR-Bench [23], take a step further by integrating both skills into a unified evaluation setting, where models must leverage strong understanding and generation capabilities to succeed. However, despite these advances, there remains a lack of a dedicated evaluation framework to measure whether UMMs truly achieve a reciprocal fusion of understanding and generation, rather than merely combining both functionalities at an engineering level. To this end, we propose **GAPEVAL**, a high-quality bidirectional benchmark specifically designed for quantifying the inherent gap of different capabilities in UMMs.

B. Detailed Metric Implementation (MIRT-MAP Version)

In this section, we present the detailed formulation of our multidimensional IRT-based capability gap quantification metric, where both text understanding and image generation abilities are jointly estimated under a Bayesian maximum a posteriori (MAP) framework.

B.1. Data Preparation

Given the evaluation results from Stage I, we first aggregate binary correctness into count statistics for each model.

Input Data Structure. For each model m_i ($i = 1, \dots, N$), we collect four counts:

- $n_i^{T\checkmark I\times}$: Text correct, Image incorrect
- $n_i^{T\times I\checkmark}$: Text incorrect, Image correct
- $n_i^{T\checkmark I\checkmark}$: Both correct
- $n_i^{T\times I\times}$: Both incorrect

From these, we derive marginal counts:

$$n_i^{\text{text-success}} = n_i^{T\checkmark I\times} + n_i^{T\checkmark I\checkmark} \quad (2)$$

$$n_i^{\text{text-fail}} = n_i^{T\times I\checkmark} + n_i^{T\times I\times} \quad (3)$$

$$n_i^{\text{image-success}} = n_i^{T\times I\checkmark} + n_i^{T\checkmark I\checkmark} \quad (4)$$

$$n_i^{\text{image-fail}} = n_i^{T\times I\times} + n_i^{T\checkmark I\times} \quad (5)$$

B.2. Multidimensional IRT Formulation

B.2.1. Model Specification

We extend the Rasch model [36] into a two-dimensional IRT framework. Each model m_i is associated with a latent ability vector:

$$\theta_i = \begin{bmatrix} \theta_i^{\text{text}} \\ \theta_i^{\text{image}} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_{\text{text}} \\ \beta_{\text{image}} \end{bmatrix}. \quad (6)$$

The success probabilities follow:

$$P_i^{\text{text}} = \frac{1}{1 + \exp(-(\theta_i^{\text{text}} - \beta_{\text{text}}))}, \quad (7)$$

$$P_i^{\text{image}} = \frac{1}{1 + \exp(-(\theta_i^{\text{image}} - \beta_{\text{image}}))}. \quad (8)$$

B.2.2. Log-Likelihood and Prior

The joint log-likelihood for all models is:

$$\mathcal{L}(\Theta, \beta) = \sum_{i=1}^N \left[n_i^{\text{text-success}} \log P_i^{\text{text}} + n_i^{\text{text-fail}} \log(1 - P_i^{\text{text}}) + n_i^{\text{image-success}} \log P_i^{\text{image}} + n_i^{\text{image-fail}} \log(1 - P_i^{\text{image}}) \right], \quad (9)$$

where $\Theta = \{\theta_1, \dots, \theta_N\}$.

To couple the two modalities, we impose a shared multivariate Gaussian prior over θ_i :

$$p(\theta_i) = \mathcal{N}(\theta_i | \mu, \Sigma), \quad \Sigma = LL^\top, \quad (10)$$

where μ is the mean vector, and L is the Cholesky factor of the covariance matrix Σ .

The total MAP objective is therefore:

$$\mathcal{L}_{\text{MAP}} = \mathcal{L}(\Theta, \beta) - \frac{1}{2} \sum_{i=1}^N (\theta_i - \mu)^\top \Sigma^{-1} (\theta_i - \mu) - \frac{N}{2} \log |\Sigma|. \quad (11)$$

This formulation allows the covariance Σ to be learned adaptively, capturing both modality-specific difficulty and cross-modality correlation.

B.3. Parameter Estimation

We optimize $\{\theta_i\}_{i=1}^N, \beta, \mu$, and L jointly via gradient-based MAP estimation (e.g., Adam). The Cholesky factorization ensures Σ remains positive definite throughout training.

B.4. Capability Gap and Normalization

After optimization, we compute for each model:

$$\Delta\theta_i = \theta_i^{\text{text}} - \theta_i^{\text{image}}, \quad (12)$$

$$\mathcal{G}_{\text{abs}}(\Delta\theta_i) = \frac{|\Delta\theta_i|}{1 + |\Delta\theta_i|}. \quad (13)$$

Here $\mathcal{G}_{\text{abs}}(\Delta\theta_i) \in [0, 1)$ represents the normalized absolute capability gap. We further apply a sigmoid normalization to each dimension:

$$\theta_{i,\text{norm}}^{\text{text}} = \sigma(\theta_i^{\text{text}}) = \frac{1}{1 + e^{-\theta_i^{\text{text}}}}, \quad (14)$$

$$\theta_{i,\text{norm}}^{\text{image}} = \sigma(\theta_i^{\text{image}}) = \frac{1}{1 + e^{-\theta_i^{\text{image}}}}, \quad (15)$$

which maps ability estimates into $(0, 1)$ for interpretability.

Post-hoc reward-penalty on the gap. To encourage consistency when the model is capable and to penalize simultaneous failures, we apply a smooth logit-space adjustment to the normalized gap. Instead of using model probabilities, we use observed co-occurrence statistics: let c_i^{SS} and c_i^{FF} be

the observed counts (or proportions) of co-success and co-failure for task i , with n_i the corresponding total number of paired observations. We define the empirical rates

$$\text{logit}(\tilde{\mathcal{G}}_i) = \text{logit}(\mathcal{G}_{\text{abs}}(\Delta\theta_i)) + (\lambda_{\text{fail}} f_i - \lambda_{\text{succ}} s_i) \quad (16)$$

so that positive λ_{fail} enlarges the gap under co-failure (penalty) and positive λ_{succ} shrinks it under co-success (reward). Unless otherwise stated, we use the default setting $\lambda_{\text{fail}} = \lambda_{\text{succ}} = 2$.

C. Reliability Analysis

To further analyze the model preference, we repeatedly calculate the Gap Scores using Gemini3-Flash. As shown in Tab. 3, the results from GPT5-mini and Gemini3-Flash share consistent relative ranking. Specifically, GPT-Image-1 exhibits the lowest gap score across different judge models. Judge models fail to exhibit a preference for the same-family model (for Gemini2.5-Flash-Image, 62.65 by Gemini3-Flash and 62.91 by GPT5-mini). The Pearson Correlation between the two judge models is 0.9656, highlighting the significant correlation.

Crucially, since most questions in **GAPEVAL** focus on the correctness of semantics (e.g., specific objects, theme), the evaluation targets are objective and definite, making it suitable for MLLM-as-a-Judge. This factual nature minimizes the room for model-specific preference bias compared to open-ended generation tasks.

Judge Model	Bagel	OneCAT	UniWorld-V1	UniPic2
Gemini3-Flash	72.96	60.67	66.64	73.39
GPT5-mini	71.52	54.73	63.47	68.97
Judge Model	Show-o2	OmniGen2	Gemini2.5-F-1	GPT-Image-1
Gemini3-Flash	72.33	88.87	62.65	47.47
GPT5-mini	67.83	89.74	62.91	50.61

Table 3. Gap Score Results from Different Judge Models.

D. Contribution

D.1. Contribution

We argue that *Synergy* **relies on Alignment**. Unlike black-box synergy evaluations, which mix capabilities, our decoupled design offers a **visible and interpretable diagnosis** of intrinsic modality gaps. This makes the evaluation **controllable and clear**, serving as a fundamental step to improve complex synergy.

D.2. Gap Score & Synergy Effects

The field of UMMs emphasizes the significance of synergy effects between two modalities. In this section, we analyze the relationship between Gap Score and Synergy Effects. We plot our Gap Score against performance on Synergy Benchmarks (GIR-Bench[23]). As shown in Fig. 7, we observe a

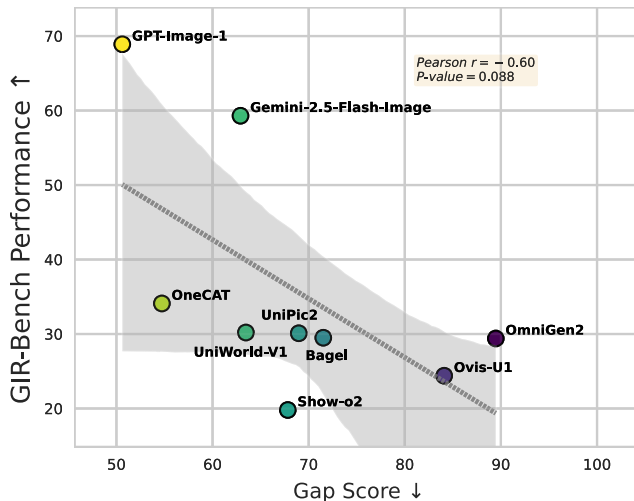


Figure 7. Correlation between Gap Score and Performance on GIR-Bench.

strong negative correlation. Models with lower Gap Scores consistently achieve higher synergy performance, mathematically confirming that narrowing the gap is a prerequisite for real-world applications.

Synergy essentially demands the complementarity of capabilities, where understanding guides generation, and generation reflects understanding. However, misalignment acts as a semantic barrier to this complementarity. If the internal representations for understanding and generation are disjoint, the information cannot be effectively transferred (*e.g.*, an editing instruction correctly understood by the encoder fails to trigger the correct generation in the decoder). Therefore, alignment is the structural prerequisite for Synergy. Our decoupled diagnosis pinpoints such gaps, providing a controllable lever for improvement.

E. Data Construction Details

E.1. Benchmark Data Construction

World Knowledge. We curate and rewrite knowledge entities from the public web and open-source datasets. For each entity, we ensure appropriate difficulty and that the target outcome can be expressed through both text and image. We author paired prompts for understanding and generation, and apply a strict cross-checking workflow to continually refine entities and prompts.

Numerical Perception. We select target object sets such as animals and vehicles, and ensure that instances of the same category appear within a single image. We use an automatic data engine to generate corresponding prompts. Images for the counting task are generated with Qwen-Image [57]. Annotators screen image quality and discard low-quality samples. For images whose object quantities do not match the targets, we rewrite the prompts and questions so that the final

items precisely enforce the counting requirement.

Instruction Following. We mine items with substantial visual changes from open-source datasets and manually construct rule-based editing prompts. For textual outputs, we guide models to describe the edited image and explicitly identify the modified regions. All items undergo strict cross-checking to ensure quality and factual correctness.

Reasoning. The reasoning set contains three subsets collected through complementary pipelines. First, the real-world reasoning covers the understanding of physical phenomena and the generation of schematic illustrations. All prompts, reference texts, and reference images are manually written and drawn with consistent visual style. Second, multiple choice reasoning is built by rendering textual problems from MMLU [51] and MMMU [68] into images. We also design algorithmic reasoning items (*e.g.*, binary tree preorder traversal and topological sorting) and produce accompanying diagrams. Each problem has a unique answer and a consistent solution key, with reference images highlighted within the answer options. Third, the image state transition reasoning task is collected from the web and open-source sources. Given several images, the task asks the model to infer and render the next state while remaining faithful to the observed evidence, which poses a higher level of difficulty. We conduct cross-review and remove items that are excessively difficult.

E.2. Empirical Study Data Collection

To rigorously investigate the behavior of Unified Multimodal Models (UMMs) under knowledge manipulation, we constructed two distinct datasets tailored for Knowledge Injection (introducing novel concepts) and Knowledge Editing (altering existing conceptual associations). The data collection and construction process is detailed below.

Object Selection and Image Collection. For the *Knowledge Injection* task, we aimed to identify entities that are absent from the models’ pre-training data. We initially curated a candidate pool of approximately 2,000 entities spanning diverse domains, including landmarks, celebrities, and biological organisms. These entities were specifically selected to be visually distinctive yet possess low public prominence (*i.e.*, long-tail distribution). For each candidate, we collected over ten high-quality images from the web.

To ensure these entities were truly “unknown” to the models, we implemented a strict filtration protocol. For each object, we randomly sampled five images to query the candidate UMMs for identification (Visual Question Answering). Simultaneously, we prompted the models with the entity names to assess their ability to generate corresponding visual representations (Text-to-Image). Only entities that all subject models failed to correctly recognize or generate were retained. This process yielded a final set of approximately

100 “unknown” objects.

For the *Knowledge Editing* task, we selected approximately 50 pairs of objects from the GenEval dataset [14]. These pairs were chosen based on conceptual relatedness or domain similarity (e.g., *boat* ↔ *car*, *camera* ↔ *microwave*) to serve as targets for conceptual swapping. To ensure high visual quality and consistency, we utilized the Qwen-Image [58] model to generate multiple images for these objects. Given that these objects are commonplace and current UMMs exhibit high performance on GenEval, we proceed with the premise that the models possess robust prior knowledge of these entities, making them suitable candidates for editing.

Training Data Construction. To facilitate the unified training of understanding and generation capabilities, we constructed a bidirectional dataset comprising both Visual Question Answering (VQA) and Text-to-Image (T2I) samples for each entity.

- **Knowledge Injection Data:** For the VQA component, we designed two types of questions for each object: open-ended queries (e.g., “*What is this object?*”) and multiple-choice questions. We constructed 5 distinct samples for each question type per object. For the T2I component, we created 10 caption-image pairs per object, consisting of 5 detailed descriptions and 5 concise captions.
- **Knowledge Editing Data:** The structure of the editing dataset mirrors that of the injection dataset but employs a counter-factual label assignment to induce knowledge swapping. Specifically, for a given object pair (Object A, Object B), the training data maps the visual representation of Object A to the textual label of Object B, and vice versa. Consequently, the VQA ground truth for an image of Object A is defined as “Object B,” and the T2I target for the prompt “Object A” is an image of Object B. This formulation forces the model to overwrite its internal alignment between the visual and textual modalities.

These strictly filtered and bidirectionally constructed datasets serve as a robust foundation for analyzing the mechanism of knowledge manipulation within UMMs.

F. Experiment Details

F.1. Benchmark Evaluation Details

In this study, we conduct a comprehensive evaluation on **GAPEVAL** using a diverse set of models, including nine unified models, four understanding-only models, and two generation-only models. Each query in **GAPEVAL** is bidirectional, capable of being answered via both textual and visual modalities. We perform ten independent sampling runs for each output modality of each question per model. We then report the average accuracy and Gap Score as the final results. Our evaluation metric comprises two primary dimensions:

The first dimension adopts the LLM-as-a-judge to assess the correctness of each response. Given that our benchmark spans four distinct categories across two modalities, we design eight specific evaluation prompts to ensure robust assessment. The prompts used for evaluation are provided in [Appendix G](#).

The second dimension focuses on the Gap Score computation. We aggregate the model performance metrics (including accuracy and other indicators) on **GAPEVAL** and leverage Multidimensional Item Response Theory (MIRT) to quantify the Gap Score. Further details regarding the Gap Score formulation are provided in [Appendix B](#).

F.2. Empirical Study Experiment Details

F.2.1. Motivation

The evaluation on **GAPEVAL** highlights a clear disparity between understanding and generation. Although UMMs can often recall embedded knowledge within one modality, they struggle to apply the same knowledge consistently across modalities. This raises a key question: **How does knowledge evolve during the training phase?**

As shown in [Table 2](#), the results quantitatively confirm that current UMMs lack genuine cross-modal consistency. To probe this issue, we conduct knowledge-oriented fine-tuning experiments by injecting or editing knowledge in one modality and evaluating its effects on both. If knowledge is shared, changes should propagate; if isolated, the effects should remain local.

The evaluation results on **GAPEVAL** reveal a clear gap between understanding and generation. While UMMs can often recall and utilize embedded knowledge to correctly answer a question in one modality, they frequently fail to produce consistent results when the same knowledge must be applied in another. This discrepancy raises a fundamental question: **Are understanding and generation truly integrated within UMMs, or do they merely coexist as separate components?**

As shown in [Table 2](#), the performance patterns observed on **GAPEVAL** provide quantitative evidence that current UMMs still struggle to achieve genuine cross-modal consistency. These findings motivate a deeper investigation into the internal organization of knowledge within UMMs: **Is the knowledge integrated, co-existing, or even conflicting across capabilities?**

To answer this question, we conduct a series of fine-tuning experiments from a knowledge-oriented perspective. Specifically, we inject or edit knowledge within one modality (understanding or generation) and then evaluate the fine-tuned model on both modalities (understanding and generation). If understanding and generation rely on a shared knowledge base, modifying knowledge on one side should lead to measurable changes on the other. Conversely, if the knowledge representations are stored separately, the impact of such

modifications will remain localized.

F.2.2. Training Strategies

Our experimental evaluation incorporates three unified models: Bagel, OmniGen2, and Show-o. For OmniGen2 and Show-o, we implement a disjoint training strategy. Under this setting, the models are fine-tuned exclusively on a single modality, either understanding or generation, to isolate specific capabilities. However, we adopt a different approach for Bagel due to constraints in its official implementation. We observe that fine-tuning Bagel on a single modality results in a complete loss of capability in the other (e.g., training solely on understanding tasks renders the model incapable of image generation). Consequently, to preserve the model’s bidirectional versatility, we employ a joint training strategy for Bagel. We integrate our knowledge manipulation dataset with the standard training data provided by the official repository to ensure robust performance across both tasks.

F.3. Metric Design

In this study, we investigate two distinct forms of knowledge manipulation: Knowledge Injection and Knowledge Editing. For understanding tasks, we utilize MLLM to verify the accuracy of the model’s textual responses.

For generation tasks, the evaluation metrics are designed to reflect the different goals of each task:

For Knowledge Editing, we employ an LLM-as-a-judge to assess the generated images. The rationale is that knowledge editing usually targets common objects (e.g., “dog”->“cat”). Since the base object is already well-known, the challenge lies in verifying semantic consistency rather than visual recognition. An LLM judge allows for a nuanced comparison between the reference and the output to confirm that the specific attributes have been edited correctly.

For Knowledge Injection, we rely on CLIP [35] score as the primary metric. This task requires the model to synthesize novel objects based on provided data. Therefore, measuring the cosine similarity between the generated image and the reference image is crucial, as it strictly penalizes the model if it fails to capture and reproduce the specific visual features of the newly injected knowledge.

The prompts used for evaluation are provided in [Table 12](#) and [Table 13](#).

G. Evaluation Prompts & Case Gallery

Prompt for Understanding on World Knowledge Task.

[Image]

[Reference_Image]

Here is the question: [Question]

Here is the answer: [Answer]

Please judge the correctness of the answer. You should follow the following rules:

1. It includes the core information present in the [Answer]
. If it does not contain the content of the reference text, judge as “not”.
 2. It reasonably describes the main subject and scene shown in the [Reference_Image].
 3. It does not need to give an exhaustive or detailed account of every feature in the image.
 4. Omissions or variations are acceptable, as long as the text covers the essential elements stated in the reference and matches the main content of the image.
 5. Only if the generated text misses the core information of the reference or fails to describe the main subject of the image should it be judged as “not”.
-

Table 4. **Prompt for Understanding on World Knowledge Task.**

Prompt for Generating on World Knowledge Task.

[Image]

[Reference_Image]

Here is the question: [Question]

Here is the answer: [Answer]

Please judge the correctness of the answer. You should follow the following rules:

1. Compare the generated image [Image] with the reference image [Reference_Image] and the caption [Question], and decide whether the image should be judged as pass (score 1) or fail (score 0).
 2. If [Image] and [Reference_Image] are identical or extremely similar (i.e., contain visual regions that look directly copy-pasted with the same pixels, appearance, texture, and details), you must judge this as plagiarism and assign score 0; this plagiarism check has the highest priority and only original, newly generated images may pass.
 3. Judge as pass if [Image] clearly presents the main subject, core scene, and key information required by [Reference_Image] and [Question]; exact reproduction of every element, attribute, arrangement, or color is not necessary, and differences in style, details, or smaller elements are allowed.
 4. Judge as fail (score 0) if [Image] misses or seriously misinterprets the core content, main objects, or key semantics described in [Reference_Image] or [Question], or if it obviously contradicts the caption or omits elements that must be strictly matched.
 5. Treat minor differences and reasonable variations as acceptable as long as the overall main information, semantics, and scene still match [Question], but never override the anti-plagiarism rule when making the final decision.
-

Table 5. **Prompt for Generating on World Knowledge Task.**

Prompt for Understanding on Reasoning Task.

[Image]

[Reference_Image]

Here is the question: [Question]

Here is the answer: [Answer]

Please judge the correctness of the answer. You should follow the following rules:

1. You are given a reasoning problem [Question], an authoritative reference_answer (the expected result or phenomenon), and a model-generated answer [Answer]. Your goal is to determine whether the final outcome/result expressed in [Answer] matches the reference_answer.
 2. Check only whether [Answer] actually provides a final result/answer to the problem; ignore any reasoning, formulas, mechanisms, or intermediate steps when making the judgment.
 3. Treat differences in wording, phrasing, or format between [Answer] and the reference_answer as acceptable, as long as they clearly describe the same final physical outcome or phenomenon.
 4. If the final result in [Answer] is present and matches the reference_answer, mark it as correct (score = 1), even if the explanation, derivation, or mechanism is incomplete or physically incorrect.
 5. If the final result in [Answer] contradicts, omits, or fails to provide the expected outcome described by the reference_answer, mark it as incorrect (score = 0).
-

Table 6. **Prompt for Understanding on Reasoning Task.**

Prompt for Generating on Reasoning Task.

[Image]

[Reference_Image]

Here is the question: [Question]

Here is the answer: [Answer]

Please judge the correctness of the answer. You should follow the following rules:

1. Compare [Image] with [Reference_Image] and [Answer] and check whether all required answer-relevant elements are present: the main result, key objects, and core information needed to visually answer the physics question posed by the problem. Major answer-relevant objects must not be missing or clearly misrepresented.
 2. Examine each main object and its physical configuration: position, alignment, orientation, relative height, order, distance, contact, and any changes such as addition, removal, joining, splitting, or shape transformation. Verify that every expected answer-relevant object from [Reference_Image] is properly accounted for in [Image].
 3. Check the physical relationships and processes: connections, supports, flows, force directions, movements and events. The depiction in [Image] must be logically and physically plausible and reflect the transformation or event described in [Answer]. Any critical new object that changes the expected physical outcome should cause rejection.
 4. Accept minor differences in style, color, artistic rendering, and irrelevant extra objects that do not change the physical result. Focus on whether the main result and key scientific meaning match [Answer] and whether the similarity between [Image] and [Reference_Image] is correct in terms of physics outcome, not in minor visual detail.
 5. Provide reasoning that clearly explains matches and differences for the above aspects. Assign “yes” (score = 1) only if the main physical result and all crucial answer elements match [Answer]; otherwise assign “no” (score = 0).
-

Table 7. **Prompt for Generating on Reasoning Task.**

Prompt for Understanding on Numerical Perception Task.

[Image]

[Reference_Image]

Here is the question: [Question]

Here is the answer: [Answer]

Please judge the correctness of the answer. You should follow the following rules:

1. Use the JSON task in [Question] (its "objects" and "number" fields) as the authoritative specification of which object types and exact counts are required. The final confirmed result stated in [Answer] must include only those specified object types, with counts that exactly match the JSON, and must not introduce any extra or non-specified objects.
 2. Count objects as individuals only when they are clearly and unambiguously described, with object-specific physical or functional features that match the corresponding JSON class (e.g., shape, color, material, labeling, size, context). Exclude partial, ambiguous, grouped, or hybrid/fused objects from all class counts, and explicitly note any such cases in the "reason" field instead of counting them.
 3. Verify that each target object's final count in [Answer] matches the JSON exactly: no overcount, undercount, or mislabeling. Check for double counting where one object might be described multiple ways, and ensure that all required target objects are present, correctly identified, and not confused with other types.
 4. Distinguish between analysis and final result: [Answer] may discuss or analyze non-target objects while reasoning, but the final confirmed result it reports must refer only to the JSON-specified object types and their counts. Any factual inconsistency or contradiction between the descriptive content and the final numbers/types should be treated as an error.
 5. In the "reason" field, detail all findings, including how counts were derived, any ambiguities, hybrids, or errors. Assign score = 1 only if the final confirmed result in [Answer] exactly matches the object types and numbers in the JSON, with no extra objects in the claimed result; otherwise assign score = 0.
-

Table 8. Prompt for Understanding on Numerical Perception Task.

Prompt for Generating on Numerical Perception Task.

[Image]

[Reference_Image]

Here is the question: [Question]

Here is the answer: [Answer]

Please judge the correctness of the answer. You should follow the following rules:

1. Use the JSON specification in [Question] (its "objects" list and corresponding "number" fields) as the exact target: [Image] passes (score = 1) only if every specified object type appears with exactly the required quantity and class, regardless of other non-target real objects that may be present.
 2. Count only distinct, real 3D object instances that fully and clearly match the visual features of their type (shape, color, label, typical context, etc.). Do not count drawings, icons, symbolic representations, or misrepresented/fused objects; each instance in overlaps, stacks, or crowds must still be individually countable and unambiguously identifiable.
 3. Assign each counted instance to the correct class exactly once: do not double count the same item due to reflections, shadows, or repeated renderings, and treat any mislabeling (e.g., calling a notebook a dictionary) or hybrid objects (e.g., "dictionary-notebook" blends) as errors that must not contribute to any class's count.
 4. Accept variations in appearance, design, pose, perspective, or partial occlusion as long as the object's identity remains clear; exclude partial or ambiguous cases where identity is uncertain. Ensure that each counted instance is classified to the single most suitable type and that no object is counted or classified more than once.
 5. In the "reason" field, provide a concise but detailed explanation of the match/mismatch logic, documenting any missed counts, misidentifications, fused or uncountable objects, or double counting. Assign score = 1 only if all specified object types match their required quantities and classes exactly under these rules; otherwise assign score = 0.
-

Table 9. Prompt for Generating on Numerical Perception Task.

Prompt for Understanding on Instruction Following Task.

[Image]

[Reference_Image]

Here is the question: [Question]

Here is the answer: [Answer]

Please judge the correctness of the answer. You should follow the following rules:

1. From [Question], understand the rule that modifies the image scenario and the reference text that concisely describes the expected result or core feature after this rule is applied. Identify the core aspect, feature, or outcome that must appear once the rule is in effect.
 2. Read [Answer] and extract all relevant features or changes it describes. Focus only on meaning: ignore extra detail, background information, unrelated content, length, and wording differences. Allow paraphrasing, scientific equivalence, and logical inference as long as the intended meaning can be reasonably matched.
 3. Accept [Answer] as correct if it clearly or implicitly describes the core feature/result stated in the reference text, or if it shows a correct understanding and application of the core change introduced by the rule (either condition is sufficient). Minimal, direct answers are acceptable as long as the expected meaning is present.
 4. Reject [Answer] if, after considering the rule, it omits or contradicts the intended meaning of the reference text, fails to reflect a correct understanding of the rule, or provides a different or incompatible interpretation of the rule. In ambiguous cases, accept only when the expected result can still be reasonably inferred from [Answer].
 5. Scoring: assign score = 1 only if [Answer] covers the core meaning of the reference text or reasonably reflects a correct understanding and application of the rule under the above conditions; otherwise assign score = 0.
-

Table 10. **Prompt for Understanding on Instruction Following Task.**

Prompt for Generating on Instruction Following Task.

[Image]

[Reference_Image]

Here is the question: [Question]

Here is the answer: [Answer]

Please judge the correctness of the answer. You should follow the following rules:

1. From [Question], fully understand the rule's intent and logic, and how it is supposed to modify the original image (objects, features, or arrangements in the original scenario). Identify the core effect or result that must appear after the rule is applied.
 2. Evaluate [Image] as the result after applying the rule to the original image: check whether all main features and modifications demanded by the rule are present, and whether any required objects or features have been unintentionally omitted. Focus on whether the rule's core meaning and result are clearly implemented, regardless of color, layout, style, or minor details.
 3. Use [Reference_Image] only as a sanity-check for the expected outcome: it is not the only correct solution and should not be used to enforce aesthetic, spatial, or stylistic accuracy. Ignore differences in object position, artistic style, decoration, or other aspects that do not directly relate to the rule's modification.
 4. Accept any plausible depiction as correct ("yes") if [Image] clearly implements the rule's effect on the original image and represents the required meaning/result, even when style or layout differ from [Reference_Image]. Reject ("no") if [Image] fails to implement the rule, omits required modifications, contradicts the rule's meaning, or shows a critical misunderstanding of the rule.
 5. In the "reason" field, clearly explain your judgment logic, focusing on how the rule was or was not correctly applied to the original image and whether the final modification in [Image] matches the intended effect of the rule.
-

Table 11. **Prompt for Generating on Instruction Following Task.**

Prompt for evaluation of und task in edit and inject knowledge.

[Image]

Here is the question: [Question]

Here is the answer: [Answer]

Please judge the correctness of the answer. You should follow the following rules:

1. Ensure the subject described in the [Answer] matches the subject in the ground_truth (whether it's an animal, object, person, etc.).
 2. If the output_text and ground_truth both describe the basic features, position, state, or other relevant characteristics of the subject consistently, it is considered correct.
 3. If there are differences in non-essential details (such as posture, angle, or state), these can be ignored, and it is still considered correct.
 4. Only when the subject described in the output_text is entirely wrong (e.g., "cat" is described as "dog") should it be considered incorrect.
-

Table 12. **Prompt for evaluation of und task in edit and inject knowledge.**

Prompt for evaluation of gen task in edit knowledge.

[Image]

Here is the question: [Question]


Here is the answer: [Answer]

Please judge the correctness of the answer. You should follow the following rules:

1. Ensure the subject depicted in the [Image] is the same as the subject in the ground_truth (whether it's an animal, object, person, etc.).
 2. If the [Image] clearly depicts the same main subject as the ground_truth, even if there are variations in its state, expression, angle, or other minor details, it is considered correct.
 3. If the output_image is chaotic, unclear, or does not represent the subject described in the ground_truth at all, it will be considered incorrect.
 4. Minor differences in non-essential features like mood, position, or posture are acceptable, as long as the subject is still clearly the same.
-

Table 13. **Prompt for evaluation of gen task in edit knowledge.**

Instruction Following

Image 	Understanding Describe the picture after a fruit knife is removed from it.
	Generation Generate an image with no fruit knife.


Reference Answer

Reference Image 	Reference Text There are no more fruit knives on the picture anymore.
---	---

Correct Understanding Answer

Model Output Okay, here's the image with the fruit knife removed. And here's a description of the modified image: The image now shows a wooden cutting board on a dark speckled countertop. On the cutting board, there is a green apple, a red apple, and two small green limes. A black spoon rests near the top of the cutting board, and another black utensil is partially visible at the bottom right. In the background, there's a spice rack with several jars and other kitchen items. The green fruit knife that was previously standing on the left side of the cutting board is no longer present.	Judgement The generated text explicitly states that the green fruit knife previously on the left of the cutting board is no longer present and describes the scene without the knife. This directly matches the reference statement that there are no fruit knives in the picture, so it reflects the rule correctly. Result True
---	--

Correct Generation Answer

Model Output 	Judgement The original image shows a small green-handled fruit knife on the cutting board. The generated image no longer contains that knife while retaining the other objects (apple, limes, spoon, cutting board). The core edit required by the rule — removal of the fruit knife — has been applied, so the modification is meaningful and correct.
Result True	

Wrong Understanding Answer

Model Output The picture shows a cutting board with a green apple, a red apple, and two limes on it. There is a green-handled knife and a black spoon on the board. The background includes a spice rack with various spices and a green spatula.	Judgement The generated text mentions a green-handled knife on the cutting board, which contradicts or at least fails to confirm the reference statement that there are no fruit knives in the picture. It does not explicitly state that the fruit knife was removed nor imply the absence of any fruit knife, so it does not reflect the required change. Result False
---	---

Wrong Generation Answer



Model Output 	Judgement The rule required removing the fruit knife from the original kitchen scene. The generated image does not show the original scene with the knife removed; instead it contains a large isolated knife on a plain background (the knife is still present and even more prominent) and the original countertop/board context is lost. Therefore the intended modification (removing the fruit knife from the original image) was not implemented.
Result True	

Figure 8. Case of Instruction Following

World Knowledge

<p>Understanding</p> <p>What animal is commonly used in cancer research because of its genetic similarity to humans?</p> <p>Generation</p> <p>Generate an image of the animal commonly used in cancer research due to its genetic similarity to humans.</p> <p>Reference</p> <p>mouse</p>	<p>Reference Image</p> 
--	---


Correct Understanding Answer

<p>Model Output</p> <p>The animal commonly used in cancer research due to its genetic similarity to humans is the mouse (and also rats).\n\nMice are particularly valuable because:\n* Their genetic makeup is surprisingly similar to humans, with many genes having a direct human counterpart.\n* They reproduce quickly and have a short lifespan, allowing researchers to study the effects of treatments and disease progression over generations in a relatively short period.\n* They are small and relatively inexpensive to house and care for in a laboratory setting.\n* Scientists can easily manipulate their genes to create models that mimic human diseases, including various types of cancer.</p> <p>Judgement</p> <p>Generated text explicitly identifies and describes a mouse, matching the reference label and the main subject visible in the image.</p> <p>Result <input type="text" value="True"/></p>	
--	--

Wrong Understanding Answer

<p>Model Output</p> <p>Rats are commonly used in cancer research due to their genetic similarity to humans. They share many of the same genes and physiological processes, making them valuable models for studying cancer and developing treatments.</p> <p>Judgement</p> <p>Generated text discusses rats and cancer research, while the reference text and image indicate a mouse. The core subject 'mouse' is missing/mismatched, so it fails to cover the reference content.</p> <p>Result <input type="text" value="False"/></p>	
---	--

Correct Generation Answer

<p>Model Output</p> 	<p>Judgement</p> <p>The generated image clearly depicts a mouse as the main subject, matching the caption. Background and composition differ from the reference but the core content (a mouse) is present.</p> <p>Result</p> <p><input type="text" value="True"/></p>
---	---

Wrong Generation Answer


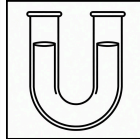
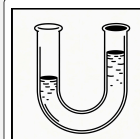
<p>Model Output</p> 	<p>Judgement</p> <p>The generated image shows a stylized rabbit/rodent hybrid with large ears and red forelimbs, not a clear mouse like the reference. The core subject 'mouse' is not faithfully represented.</p> <p>Result <input type="text" value="False"/></p>
--	---

Figure 9. Case of World Knowledge

Reasoning

<p>Image</p> 	<p>Understanding</p> <p>What happens to the water level in a U-shaped tube with a balanced water surface, connected to atmospheric pressure, when a wind blows over the right side of the tube, causing the air pressure to drop?</p>	<p>Generation</p> <p>Generate an image of a minimalist black-and-white line drawing depicting a U-shaped tube with a balanced water surface initially connected to atmospheric pressure, where a wind blowing over the right side of the tube causes a drop in air pressure, resulting in the water level rising on the right side of the tube due to the imbalance, illustrated without any people or text, in a style similar to the uploaded images.</p>
---	--	--

Reference Answer

<p>Reference Image</p> 	<p>Reference Text</p> <p>As the air pressure on the right side drops due to the wind, the water level on the right side rises, creating an imbalance between the two sides of the U-tube.</p>
---	--

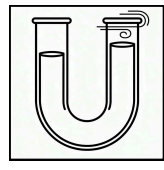
Correct Understanding Answer

<p>Model Output</p> <p>Okay, here's the image with the fruit knife removed. And here's a description of the modified image:</p> <p>The image now shows a wooden cutting board on a dark speckled countertop. On the cutting board, there is a green apple, a red apple, and two small green limes. A black spoon rests near the top of the cutting board, and another black utensil is partially visible at the bottom right. In the background, there's a spice rack with several jars and other kitchen items. The green fruit knife that was previously standing on the left side of the cutting board is no longer present.</p>	<p>Judgement</p> <p>When wind blows over the right side of the U-shaped tube, the air pressure on that side drops due to Bernoulli's principle. This lower pressure on the right side will cause the water level on the right side to rise, and consequently, the water level on the left side will fall. The system will try to equalize the pressure difference.</p> <p>Result</p> <p>True</p>
--	--

Wrong Understanding Answer

<p>Model Output</p> <p>user\nOriginal image: A U-shaped tube with a balanced water surface, connected to atmospheric pressure. Now, a wind blows over the right side of the tube, causing the air pressure to drop. What happens to the water level?\nassistant\nWhen wind blows over the right side of the U-shaped tube, it causes a decrease in air pressure on that side due to the Bernoulli effect. This decrease in pressure results in a higher pressure on the left side of the tube relative to the right side.\nAs a result, the liquid (water) will move from the area of higher pressure (left side) to the area of lower pressure (right side). The water level on the right side of the tube will rise, while the water level on the left side will fall. This movement continues until the pressure difference is balanced again, which typically means the water levels will eventually equalize.</p>	<p>Judgement</p> <p>The generated text correctly states the right-side pressure drop causes the right water level to rise and the left to fall, matching the reference; however it then incorrectly claims the water levels will 'eventually equalize,' which contradicts the expected sustained imbalance. Because of this contradiction, the final result does not match.</p> <p>Result</p> <p>False</p>
---	--

Correct Generation Answer

<p>Model Output</p>  <p>Result</p> <p>True</p>	<p>Judgement</p> <p>Plagiarism check: The generated_image is not a direct copy of the reference_image (same general U-tube style but different internal rendering and added wind glyphs). It therefore does NOT trigger the PLAGIARISM rule.\nDetailed checklist comparison and reasoning (points 1-20):\n1) Does generated_image include all required answer elements from reference_text?\n - Yes. The reference_text says the right-side water level should rise (right column higher than left). The generated_image clearly show the right side higher; the liquid surfaces appear essentially equal in height.\n2) Does it visually answer the question posed by problem_image and problem_text?\n - Yes. The problem asks what happens when pressure on the right drops; [...detailed description...] \nFinal verdict: score = 1 (the generated_image visually and accurately reflect the main answer/result).</p>
---	--

Wrong Generation Answer

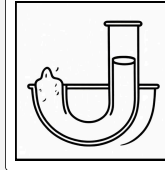
<p>Model Output</p>  <p>Result</p> <p>True</p>	<p>Judgement</p> <p>The rule required removing the fruit knife from the original kitchen scene. The generated image does not show the original scene with the knife removed; instead it contains a large isolated knife on a plain background (the knife is still present and even more prominent) and the original countertop/board context is lost. Therefore the intended modification (removing the fruit knife from the original image) was not implemented.</p>
--	--

Figure 10. Case of Reasoning