AEGIS: Exploring the Limit of World Knowledge Capabilities for Unified Multimodal Models

Supplementary Material

In Sec. A, we illustrate the details of covered topic type in AEGIS. In Sec. B, we explain the details of each covered reasoning type in AEGIS. In Sec. C, we conduct more in-depth analysis to reveal the key results in world knowledge evaluation. And in Sec. D, we illustrate the essential prompts used in AEGIS.

A. AEGIS Topic Type Descriptions

AEGIS organizes real-world knowledge into three domains (STEM, Humanities, Daily Life) and subdivides each topic into finer sub-topics to assess complementary facets.

A.1. STEM

The STEM topic assesses proficiency in Science, Technology, Engineering, and Mathematics, focusing on quantitative reasoning, application of physical and mathematical principles, and problem solving grounded in formal methods. It includes:

- **Biology** assesses knowledge related to biological common sense, including representative species and ecological traits, fundamental life processes, and biological concepts that carry cultural relevance. *Example: Please draw a picture of a female modern relative in Asia of the animal in* `<image1>`.
- **Chemistry** focuses on chemical substances, everyday chemical phenomena, and chemistry embedded in traditional crafts, covering common substances' uses, safety awareness, and widely known processes across cultures and industries. *Example: Given the two compounds shown in* `<image1>` *and* `<image2>`*, what is the expected reaction product in concentrated sulfuric acid?*
- **Mathematics** evaluates understanding of foundational mathematical concepts, common geometric figures, and everyday applications, emphasizing arithmetic, measurement, and shape recognition commonly taught across cultures. *Example: The image* `<image1>` *shown represents a cubic function. If the coefficient of the cubic term is 1/2, what is the coefficient of the linear term?*
- **Medicine** examines basic medical and health literacy, including disease prevention, first aid fundamentals, and concepts or tools widely recognized in both traditional and modern healthcare practices. *Example: How does salicylic acid* `<image1>` *enhance therapeutic efficacy? please draw the Chemical bond-line formula of the improved drug.*
- **Physics** focuses on everyday physical phenomena and foundational concepts—mechanics, thermodynamics, electromagnetism, optics, and acoustics—highlighting intuitive, real-world applications and explanations. *Example: According to the principle of thin-film interference, please color the blank areas in the diagram* `<image1>`.
- **Astronomy & Geography** assesses recognition of typical celestial and geographic features, including naked-eye sky phenomena, seasonal and directional knowledge, and culturally emblematic landmarks and biomes. *Example:* `<image1>` *shows what a location at 60 degrees north latitude looked like before 1908. Please draw what the same location looked like after 1908.*
- **IT** focuses on common digital literacy and information technology concepts, including basic computing and networking, routine data and security practices, and widely used software/hardware terms. *Example:* `<image1>` *shows a diagram of the CPU architecture. Please use the same color scheme to draw a diagram of the architecture of another common computing chip.*

A.2. Humanities

The Humanities topic evaluates understanding of human society, culture, and creative expression, emphasizing interpretive reasoning, historical analysis, contextual understanding, and critical evaluation of artifacts and practices. It includes:

- **Agriculture** evaluates knowledge of agricultural practices, crops, tools, and food systems across regions, including traditional and modern methods and their cultural-economic significance. *Example: Using the label provided in the image* `<image1>`*, please colour the map* `<image2>` *according to the proportion of hybrid rice cultivated relative to total rice acreage.*
- **History** examines recognition of major historical events, periods, figures, and artifacts, emphasizing chronology, causation, and cultural impact. *Example: Which event in the Qing Dynasty is similar to the one shown in this image* `<image1>`*?*
- **Movie** assesses familiarity with influential films, genres, directors, iconic scenes, and culturally significant cinematic symbols. *Example: What is the MacGuffin of the 1942 Academy Award for Best Original Screenplay?*
- **Music** focuses on musical traditions, instruments, genres, and notable composers or performers, highlighting stylistic features and cultural contexts. *Example: Generate a simple sheet music score of 'Twinkle Twinkle Little Star' in C major.*

- **Art** evaluates understanding of visual arts, styles, techniques, movements, and canonical works or artists across cultures and eras. *Example: Edit <image1> to show Mona Lisa looking away with a disdainful expression and holding up a sign indicating she doesn't want her photo taken.*
- **Culture** assesses broader cultural practices, norms, heritage items, and symbols that define collective identities and social life. *Example: Replace the outer skin of the three pastries in the middle of <image1> with the style of North China.*
- **Architecture** tests recognition of architectural styles, structural features, landmark buildings, and the historical-technological contexts of the built environment. *Example: What is another religious sightseeing location in the same city as <image1>?*

A.3. Daily Life

The Daily Life topic evaluates practical knowledge and daily reasoning in common modern contexts, emphasizing situational understanding, routine decision making, and the recognition of tools, activities and media encountered in daily environments.

- **Activity** evaluates familiarity with common daily activities and leisure practices, including their typical tools, settings, and procedural steps. *Example: Based on this screenshot <image1>, is there any Argentine player in an offside position?*
- **Anime** assesses recognition of notable anime series, characters, visual tropes, and stylistic conventions, as well as culturally salient symbols in animated media. *Example: Where did the protagonist of One Piece go after bidding farewell to the Empress and before witnessing his brother's death?*
- **Game** focuses on understanding of video and tabletop games, including iconic titles, genres, gameplay elements, and distinctive in-game artifacts or interfaces. *Example: What is another well-known game produced by the team leader of the 2022 TGA Game of the Year for Mobile?*
- **Photography** examines knowledge of photographic equipment, techniques, genres, and visual conventions used in image capture and editing workflows. *Example: Draw the photo of the girl <image1>, from the illustration to the cosplay photo from the most famous anime expo in the world. But keep the original background.*
- **Engineering** evaluates practical understanding of daily engineering artifacts, mechanisms, household devices, and basic technical operations relevant to daily environments. *Example: Add appropriate materials to <image1> to make it a simple distiller.*
- **Food** assesses recognition of ingredients, dishes, cooking methods, dining customs, and nutrition concepts commonly encountered in daily meals. *Example: Add Sichuan characteristics to this dish <image1>.*
- **Traffic** tests the ability to identify transportation modes, road signs, traffic rules, and navigation conventions used in urban mobility. *Example: Draw the fastest rail transit route from the tower location to the red dot location in <image1>.*

B. AEGIS Reasoning Type Descriptions

Beyond general world knowledge, AEGIS further probes LLM's capacity to follow obfuscated instructions by evaluating its underlying reasoning skills. Specifically, AEGIS categorizes reasoning into six types:

- **Spatial Reasoning** evaluates the ability to infer relationships involving position, distance, orientation, containment, and part–whole layout in 2D/3D space, which accounts for 10.9% of the entire benchmark. *Example: Given the front view <image1>, top view <image2>, and right side view <image3> of a 3D object, draw a picture of its isometric projection.*
- **Temporal Reasoning** assesses understanding of temporal order, duration, concurrency, and schedules, including before/after relations and timeline consistency, which accounts for 12.2% of the entire benchmark. *Example: Edit it to show how <image1> looks today.*
- **Causal Reasoning** examines the ability to identify cause–and–effect relations, necessary/sufficient conditions, and outcomes of interventions or counterfactual changes, which accounts for 12.2% of the entire benchmark. *Example: Infer a unified astronomical event based on <image1> and <image2>.*
- **Comparative Reasoning** concerns any comparison involving two or more entities along one or multiple dimensions, and drawing conclusions based on their relative differences or rankings, which accounts for 15.4% of the entire benchmark. *Example: <image1> and <image2>, which requires more cooking steps?*
- **Analogical Reasoning** evaluates mapping relational structure from a known scenario to a novel one, recognizing proportional or functional analogies, which accounts for 9.4% of the entire benchmark. *Example: Just as <image1> is to his corresponding anime work, who is the character in Naruto occupying a similar position?*
- **Logical Reasoning** emphasizes drawing conclusions that follow coherently from stated facts, rules, or constraints in everyday contexts, which accounts for 36.3% of the entire benchmark. *Example: Draw a Venn diagram with three intersecting sets A, B, and C, and shade the region corresponding to $(A \cap B) \cup C$.*

Table A. Comprehensive Performance Comparison for Gemini Nano Banana (Gemini for short) and GPT-4o with GPT-Image-1 (GPT for short) with different types of prompts and external web search tools.

Model	Understanding			Generation			Editing			Interleaved Generation			Overall
	STEM	Humanity	Life	STEM	Humanity	Life	STEM	Humanity	Life	STEM	Humanity	Life	
Gemini	64.5	65.7	55.0	42.6	49.5	45.5	44.4	62.4	54.2	50.2	41.6	43.4	52.9
Gemini w/ Web Search	64.9	69.2	60.0	-	-	-	-	-	-	-	-	-	-
Gemini w/ GPT prompt	63.8	64.7	56.2	37.8	46.0	44.3	41.0	61.4	51.3	47.3	43.6	43.3	51.1
Gemini w/ Gemini prompt	63.4	72.9	57.6	45.3	57.9	52.2	42.2	63.9	47.8	50.8	49.6	47.2	55.2
Gemini w/ Clear prompt	72.4	74.3	72.1	53.6	67.6	62.6	54.2	68.4	70.4	53.1	62.0	54.0	65.2
Gemini w/ Clear & Web	72.8	80.8	73.3	-	-	-	-	-	-	-	-	-	-
GPT	52.9	50.9	46.9	38.2	51.6	42.8	39.4	53.2	45.2	38.9	34.7	33.0	45.7
GPT w/ GPT prompt	48.4	52.0	43.4	35.6	52.8	41.9	38.1	54.9	46.0	33.5	33.6	34.7	44.7
GPT w/ Gemini prompt	57.0	66.7	53.2	42.6	57.3	47.0	39.0	59.3	45.4	44.6	36.1	41.8	50.8
GPT w/ Clear prompt	61.6	65.4	62.7	52.1	71.5	61.6	49.2	66.9	56.9	51.0	57.2	49.2	60.0
Gemini-3-Pro	77.7	79.3	70.4	62.2	64.4	58.2	64.1	67.8	58.5	42.6	40.2	38.5	64.3

C. Additional Experiments

In this section, we provide additional experiments based on Gemini Nano Banana [1] (Gemini for short) and GPT-4o with GPT-Image-1 [3] (GPT for short) to examine how prompts of varying specificity affect performance across tasks. We further disentangle common-sense knowledge from the UMMs via controlled ablations to isolate module-specific issues and quantify their impact. We also investigate the upperbound of UMMs by evaluating the state-of-the-art Gemini-3-Pro (*i.e.*, Nano Banana Pro).

C.1. Evaluation on UMM Rewritten Prompts

Beyond investigating the impact of external reasoning modules [5] on Bagel [2], we conducted an ablation study to further isolate the effect of reasoning quality. Specifically, we employed a "self-reasoning" strategy wherein the model first rewrites the raw prompt to generate a "clear prompt," thereby mitigating the need for downstream reasoning by resolving ambiguities, identifying entities, and making implicit context explicit. Surprisingly, while manually verified clear prompts generally yield substantial gains, we observed **divergent effects** with self-rewriting: prompts rewritten by GPT-4o [3] resulted in performance degradation compared to raw inputs, whereas those rewritten by Gemini Nano Banana [1] led to performance improvements. To validate this disparity, we performed a cross-model evaluation by swapping the rewritten prompts, *i.e.*, feeding GPT-4o with Gemini-rewritten prompts and vice versa. The results were consistent: GPT-generated rewrites caused performance drops across models, while Gemini-generated rewrites consistently yielded gains. These findings strongly suggest that Gemini Nano Banana possesses superior reasoning capabilities for instruction disambiguation compared to GPT-4o. Consequently, leveraging LLMs with advanced reasoning capabilities offers a promising avenue to mitigate the challenges posed by ambiguous or reasoning-intensive instructions in UMMs, thereby benefiting diverse tasks across a broad spectrum of world knowledge [4].

C.2. Gemini-3-Pro has better World Knowledge

As discussed in Sec.4 in the main paper, introducing training data with better quality and more abundant domain aspects can improve the world knowledge understanding abilities. However, those of the other tasks are not verified. Therefore, we evaluate Gemini-3-Pro, *i.e.*, the extended version of Gemini Nano Banana, on AEGIS benchmark. As shown in the bottom of Table A, even using the reasoning-enhanced prompts for inference, the overall performance of Gemini-3-Pro (64.3) is still much higher than that of Gemini (52.9), and achieves comparable performance with Gemini using clear prompts. These promising results indicate that better pretraining data benefits to the world knowledge capabilities, and also show that Gemini-3-Pro is the state-of-the-art UMM in world knowledge understanding and generation aspects.

C.3. Evaluation on Web Search

Following our investigation into disambiguating complex instructions, we further explored methods to mitigate prediction errors and hallucinations by incorporating new external knowledge. Intuitively, integrating a search engine should provide the essential, up-to-date information required for accurate responses. To assess this, we conducted an ablation study evaluating the impact of web search augmentation on model performance. Specifically, we enabled the Google Search tool for Gemini (specifically, Gemini-2.5-Flash-Image) to ground its responses in current events and verifiable web-based facts. Counterintuitively, activating web search only results in marginal performance improvement in understanding tasks. Especially, for questions in STEM topics, the performance gain is only 0.4. A plausible explanation is that while search tools effectively acquire external world knowledge—crucial for verifying facts or re-

trieving recent events—they do not inherently strengthen the model's core reasoning capabilities. This finding aligns with human problem-solving behaviors: effective solutions rarely emerge from directly querying a complex, raw problem into a search engine. Instead, successful problem solving typically necessitates an initial reasoning phase to formulate clear, targeted queries before consulting external resources. To verify our hypothesis, we further integrate the web search tool into Gemini with clear prompts, to investigate whether clearer and more effective problem description leads to more precise search results as auxiliary knowledge. As shown in Table A, by easing the problem with clear prompts, the understanding performance largely increases. Especially, the performance gain of humanity and life understanding questions are both larger than 10.0, which indicates web search tools can improve the world knowledge capabilities of UMMs with clear problem descriptions. These results also imply the significance of inherent reasoning capabilities of UMMs during inference.

C.4. Investigation into Module-Specific Bottlenecks

Furthermore, we aimed to identify which component acts as the *primary limiting factor* for world knowledge capabilities in UMMs: the LLM component or the visual decoder component. To locate the source of errors, we analyzed failure cases from Gemini. Specifically, we utilized Gemini-2.5-Pro to rewrite the original prompts via a self-reflection procedure, ensuring all implicit knowledge was made explicit. We then fed these rewritten prompts back into Gemini for generation. Fig. A presents a comparison of images generated from raw prompts, rewritten prompts, and verified clear prompts. Crucially, we observed that the LLM component successfully articulated the key visual attributes in the rewritten text (*e.g.*, correctly identifying "Hu Tao" or "Michelangelo"). However, the visual decoder failed to render these concepts consistently, deviating from both the clear-prompt outputs and the ground truth. This discrepancy suggests that the *visual decoder* restricts world knowledge capabilities in UMMs, likely due to insufficient knowledge encoding within the decoder itself or extreme sensitivity to input phrasing.

To rigorously verify this hypothesis, we conduct a follow-up experiment using extremely detailed descriptions. We use Gemini-2.5-Pro to generate comprehensive visual descriptions that explicitly outline every keypoint required for generation or editing, effectively bypassing the model's need to recall visual attributes. We then feed these descriptions into Gemini. As shown in Fig. D, despite the LLM providing highly accurate and detailed visual instructions, *only the first* example shows a plausible result (with marginal shape discrepancies), while the others remained incorrect. These results definitively verify that the visual generation module is the bottleneck, identifying a misalignment between the model's strong textual understanding and its weaker visual generation capabilities, consistent with the performance gaps observed in AEGIS.

D. Essential Templates Used in AEGIS

Finally, we provide templates used in the AEGIS dataset annotations, including the checklist generation prompt in Fig. E and evaluation prompts in Fig. F.

References

[1] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 3

[2] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 3

[3] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 3

[4] Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Kunpeng Ning, Bin Zhu, et al. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025. 3

[5] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837, 2022. 3

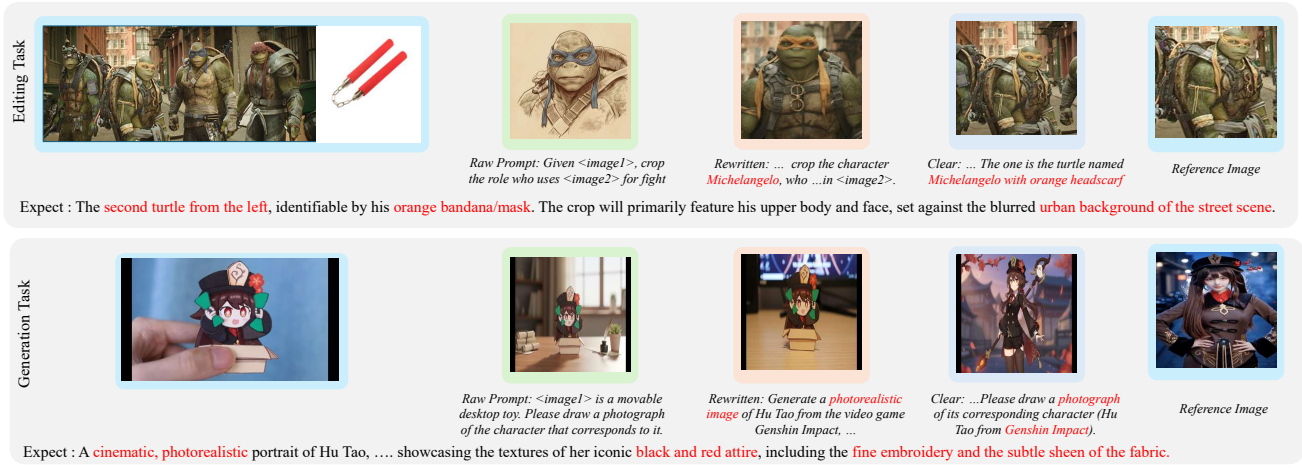


Figure A. Visualization of failure cases with raw and LLM rewritten prompts. We highlight the keypoints in the answers by red color. Though external reasoning modules (*e.g.*, Gemini) can ease the generation difficulty by rewriting complex prompts, there still exist gaps towards precise reasoning capabilities under diverse tasks across world knowledge aspects.

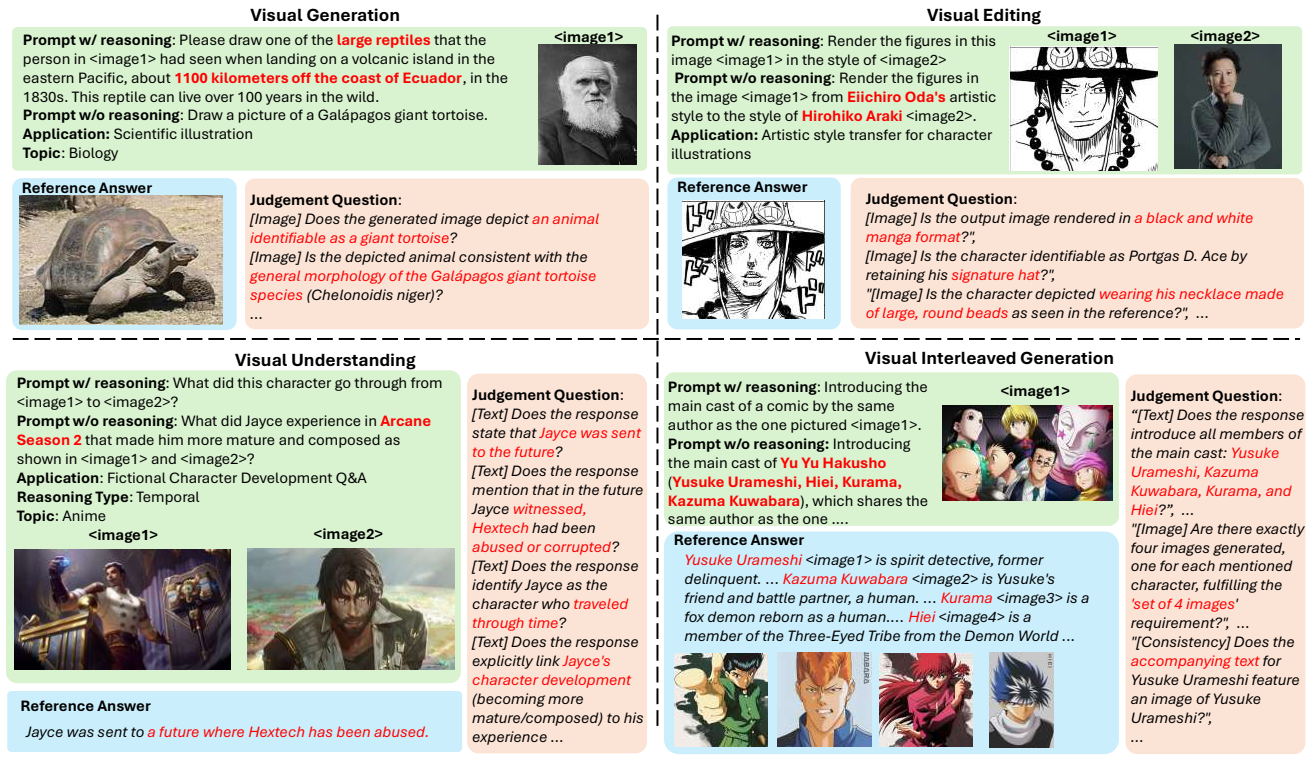


Figure B. Visualization of more questions in AEGIS benchmark.

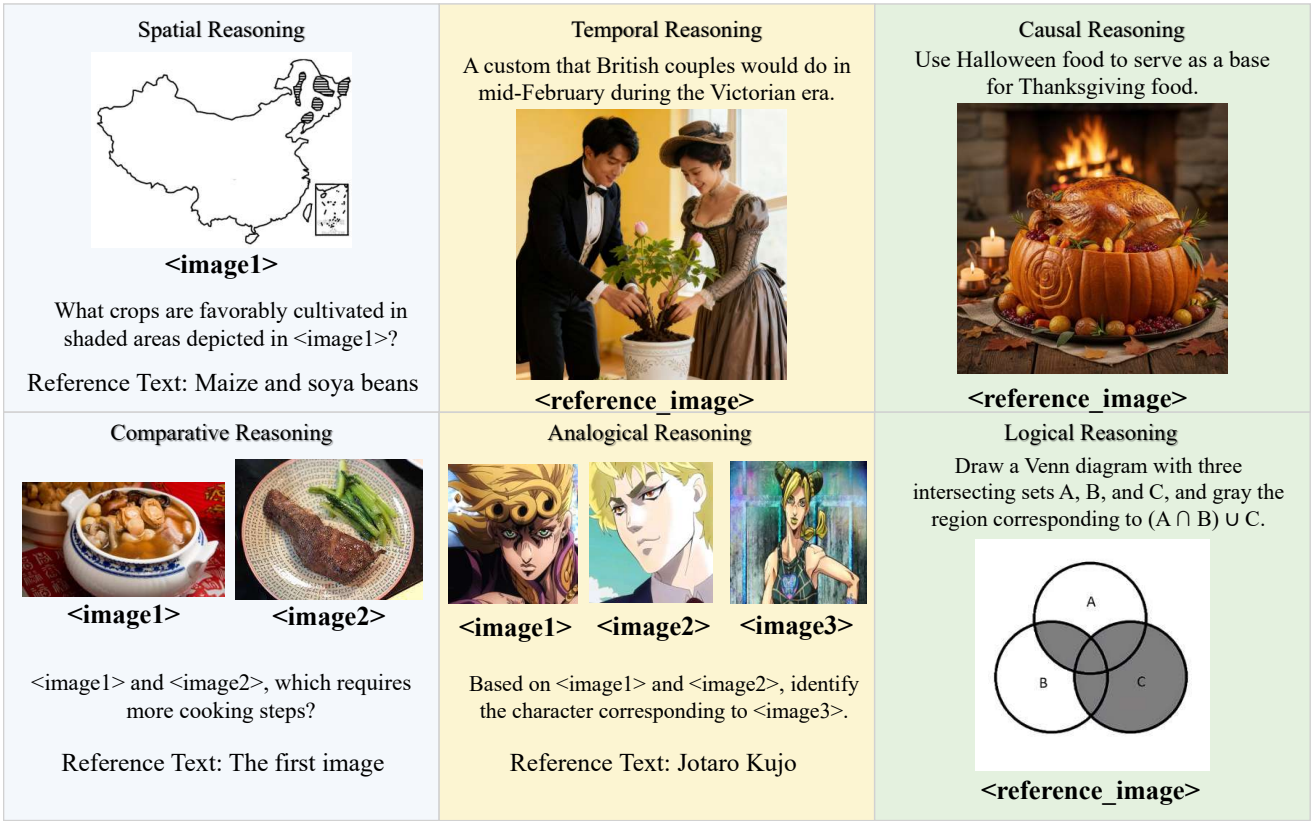
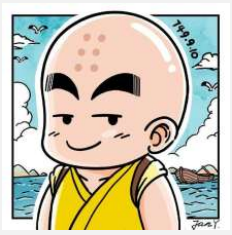


Figure C. Visualization of questions with different reasoning types. AEGIS includes various reasoning types in questions, covering common scenarios in practical applications.

<image1>



Prompt w/o reasoning: Draw a picture of a Galápagos giant tortoise where <image>'s master lives.

Reference Image

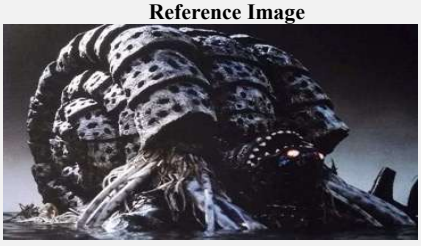
LMM Expect: A small, vibrant pink house labeled "Kame House" sits on a tropical island surrounded by crystal-clear blue water, with palm trees.

UMM Output

<image1>

Prompt w/o reasoning: Draw Gatanothor from <image1> TV series.

Reference Image



LMM Expect: A massive, grotesque creature resembling a giant shell with tentacle-like appendages, glowing red eyes, and a dark, eerie presence emerging from the water.

UMM Output

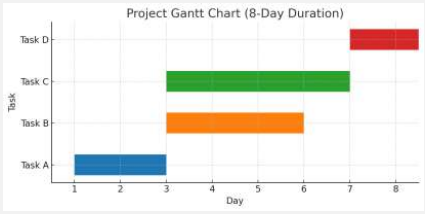
Pure Text

Prompt w/o reasoning:

Create a Gantt chart representing a project with a total duration of 8 days:

Task A: Days 1–2
Task B: Days 3–5
Task C: Days 3–6
Task D: Days 7–8

Reference Image



LMM Expect: Task A: A bar starting at Day 1 and ending at Day 2.Task B: A bar starting at Day 3 and ending at Day 5. Task C: A bar starting at Day 3 and ending at Day 6, overlapping partially with Task B. Task D: A bar starting at Day 7 and ending at Day 8.

UMM Output

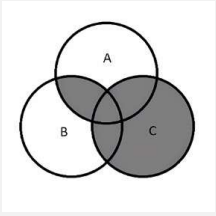
Pure Text

Prompt w/o reasoning:

Draw three overlapping circles labeled A, B, and C, with no fill color by default.

Shade the overlapping region between circles A and B in gray, and fill the entire area of circle C in gray.

Reference Image



LMM Expect: A Venn-like diagram with three overlapping outline circles labeled A, B, and C, where the A–B overlap is shaded gray and the entire circle C is filled gray.

UMM Output

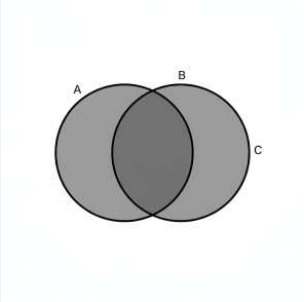


Figure D. Visualization of failure cases with refined LLM descriptions of clear prompts from Gemini. One can find that even the descriptions precisely illustrate the answers, the visual decoder still usually struggles with generating correct answers.

Task: Comprehensive Task-Aware Evaluation Question Generation

Objective:
Your task is to act as an expert evaluator. Your goal is to generate a detailed, objective, and verifiable checklist of questions to evaluate a multimodal model's output. The questions must be grounded in the provided reference materials (`Reference Text` and `Reference Images`) while ensuring they cover the core concepts defined by the `Keywords` and `Category`.

Inputs:
- **Task Type:** {task} (The category of the task, which dictates the output modality: `understanding`, `editing`, `generation`, or `interleaved`.)
- **Keywords:** {keywords} (A set of core concepts that must be addressed.)
- **Category:** {category} (The broader topic for contextual relevance.)
- **Clear Prompt:** {clear_prompt} (The original instruction given to the model being evaluated.)
- **Reference Text:** {ref_text} (The **primary source of truth** for text-based facts.)
- **Reference Images:** <ref_images_placeholder_list> (The **primary source of truth** for visual facts.)

Core Instructions:

1. **Determine Output Modality via Task Type:** This is your first and most important step. It dictates which question tags you are allowed to use.
   * **If `Task Type` is `understanding` (text-only output):** You must **ONLY** use the `[Text]` tag.
   * **If `Task Type` is `editing` or `generation` (image-only output):** You must **ONLY** use the `[Image]` tag.
   * **If `Task Type` is `interleaved` (text and image output):** You may use `[Text]`, `[Image]`, and `[Consistency]` tags. `[Consistency]` questions are crucial here.

2. **Ground Questions in Reference Material:** All questions must be derived from specific, verifiable details found in the `Reference Text` and/or `Reference Images`. Do not invent questions that cannot be answered by the reference materials.

3. **Focus Questions using Keywords and Category:** Use the `Keywords` and `Category` as a lens to focus your attention. Prioritize creating questions about the details in the reference materials that are most relevant to these keywords and the overall topic. For instance, if a keyword is "egg", generate specific questions about how the eggs are used (e.g., separated, whipped) as described in the `Reference Text`.

4. **Formulate Specific, Objective & Tagged Questions:**
   * Each question must be prefixed with a modality tag (`[Text]`, `[Image]`, `[Consistency]`).
   * Questions must be objective and factual. **AVOID** subjective assessments of quality, style, tone, theme, or artistic merit.
   * **Bad (Subjective):** `[Image]` Is the photo aesthetically pleasing?
   * **Bad (Vague):** `[Text]` Does the text talk about the keywords?
   * **Good (Specific & Objective):** `[Text]` Does the recipe state to bake the cake for 60 minutes at 150°C?

5. **Ensure Comprehensive Coverage & No Redundancy:** The final checklist should cover all critical aspects related to the keywords without asking repetitive questions about the same detail.

Output Format:
Noted that the output should be a string that can be directly converted into a Python list using the json.loads() function, and The value should be an array of strings, where each string is a tagged evaluation question.

Generate the checklist for the provided inputs now.

Figure E. Checklist generation prompts in AEGIS benchmark. We formulate the LLM-as-a-Judge evaluation by a series of atomic “Y/N” questions to avoid ambiguous judgments.

Task: Comprehensive Task-Aware Evaluation

Objective:
You are an expert for world-knowledge-based evaluation. Your task is to verify whether the output image or text meets a series of checklists and provide your reasoning for the evaluation.

INPUT FORMAT:
You will be provided with the following fields:
- **Task Type**: The category of the task, which dictates the output modality: `understanding`, `editing`, `generation`, or `interleaved`.
- **Category**: The broader topic for contextual relevance.
- **Clear Prompt**: The original instruction given to the model to output the image or text or both.
- **Reference Text**: The **primary source of truth** for text-based facts.
- **Reference Images**: The **primary source of truth** for visual facts.
- **Output Text**: The output texts needed to be evaluated by you.
- **Output Image**: The output images needed to be evaluated by you.
- **Checklist**: A series of checklists for the output. Each item in the checklist contains one of the tags: `[Text]`, `[Image]`, and `[Consistency]`.

For item with tag `[Text]`, Please focus on analyzing whether the **Output Text** meets the requirements of the checklist item. You can treat the **Reference Text** as a reference that perfectly meets all checklist items. However, the output does not have to be identical to the **Reference Text**. As long as it meets the requirements of the checklist item, it can be marked as passing (Y).
For item with tag `[Image]`, Please focus on analyzing whether the **Output Images** meets the requirements of the checklist item. You can treat the **Reference Images** as a reference that perfectly meets all checklist items. However, the output does not have to be identical to the **Reference Images**. As long as it meets the requirements of the checklist item, it can be marked as passing (Y).
For item with tag `[Consistency]`, Please focus on analyzing whether the **Output Text** and **Output Images** are consistent with each other according to the checklist item. You can treat the **Reference Text** and **Reference Images** as a reference that perfectly meets all checklist items. However, the output does not have to be identical to the **Reference Text** and **Reference Images**. As long as it meets the requirements of the checklist item, it can be marked as passing (Y).

TASK & OUTPUT REQUIREMENTS:
Your output must be a single valid JSON object. The Json object should be dict with the following keys:
- **Answer List**: A list of answers for the checklist. Each item in the answer list corresponds to an item in the checklist in order. Each entry is either "Y" or "N," representing "yes" or "no," respectively.
- **Reason List**: The reasoning for the evaluation. Each item in the Reason List explains the reason for the corresponding Y/N in the Answer List.

Evaluate the output according to all requirements above. Ensure the output is valid JSON.

Figure F. DCE Evaluation prompts in AEGIS benchmark. We predict “yes / no” judgements for all atomic judgement questions, and calculate the percentage of “yes” judgements as final scores.