UEval: A Benchmark for Unified Multimodal Generation

Bo Li Yida Yin Wenhao Chai Xingyu Fu* Zhuang Liu*

Princeton University

Website: https://zlab-princeton.github.io/UEval

Code Dataset Leaderboard

Abstract

We introduce UEval, a benchmark to evaluate *unified models*, *i.e.*, models capable of generating both images and text. UEval comprises 1,000 expert-curated questions that require both images and text in the model output, sourced from 8 real-world tasks. Our curated questions cover a wide range of reasoning types, from step-by-step guides to textbook explanations. Evaluating open-ended multimodal generation is non-trivial, as simple LLM-as-a-judge methods can miss the subtleties. Different from previous works that rely on multimodal Large Language Models (MLLMs) to rate image quality or text accuracy, we design a rubric-based scoring system in UEval. For each question, reference images and text answers are provided to a MLLM to generate an initial rubric, consisting of multiple evaluation criteria, and human experts then refine and validate these rubrics. In total, UEval contains 10,417 validated rubric criteria, enabling scalable and fine-grained automatic scoring. UEval is challenging for current unified models: GPT-5-Thinking scores only 66.4 out of 100, while the best open-source model reaches merely 49.1. We observe that reasoning models often outperform non-reasoning ones, and transferring reasoning traces from a reasoning model to a non-reasoning model significantly narrows the gap. This suggests that reasoning may be important for tasks requiring complex multimodal understanding and generation.

1 Introduction

Unified multimodal models (Tong et al., 2024b; Zhou et al., 2025a; Deng et al., 2025) aim to integrate multimodal understanding and generation capabilities within a single system. Current evaluations of these models are largely confined to two paradigms: visual question answering (Marino et al., 2019; Liu et al., 2024b; Yue et al., 2024; Fu et al., 2025), which requires generating a textual answer from one or more input images, and text-to-image generation (Huang et al., 2023; Ghosh et al., 2023; Lin et al., 2024), which takes a textual description as input and asks the model to produce a corresponding image.

These paradigms overlook a central component of multimodal reasoning scenarios: unified multimodal generation that *produces both text and images* in response to a single query (Figure 1). In many real-world tasks, effective responses require images to illustrate specific concepts while simultaneously producing text to explain those visual elements. Without such evaluation, existing benchmarks fail to capture the rich interplay between language and vision that characterizes real-world multimodal reasoning.

While recent efforts (An et al., 2024; Liu et al., 2024a; Xia et al., 2025; Niu et al., 2025; Zhao et al., 2025) have proposed new benchmarks to evaluate unified models, there remains a lack of standardized approaches for evaluating unified multimodal generation. To address this gap, we introduce UEval, a challenging benchmark to assess unified models (Wang et al., 2024; Chen et al., 2025c; Yang et al., 2025; Google, 2025b; Xie et al., 2025) at scale. Unlike prior benchmarks, UEval requires models to reason and respond to complex user queries jointly in images and natural language, providing a rigorous testbed across diverse real-world scenarios.

* Co-advising

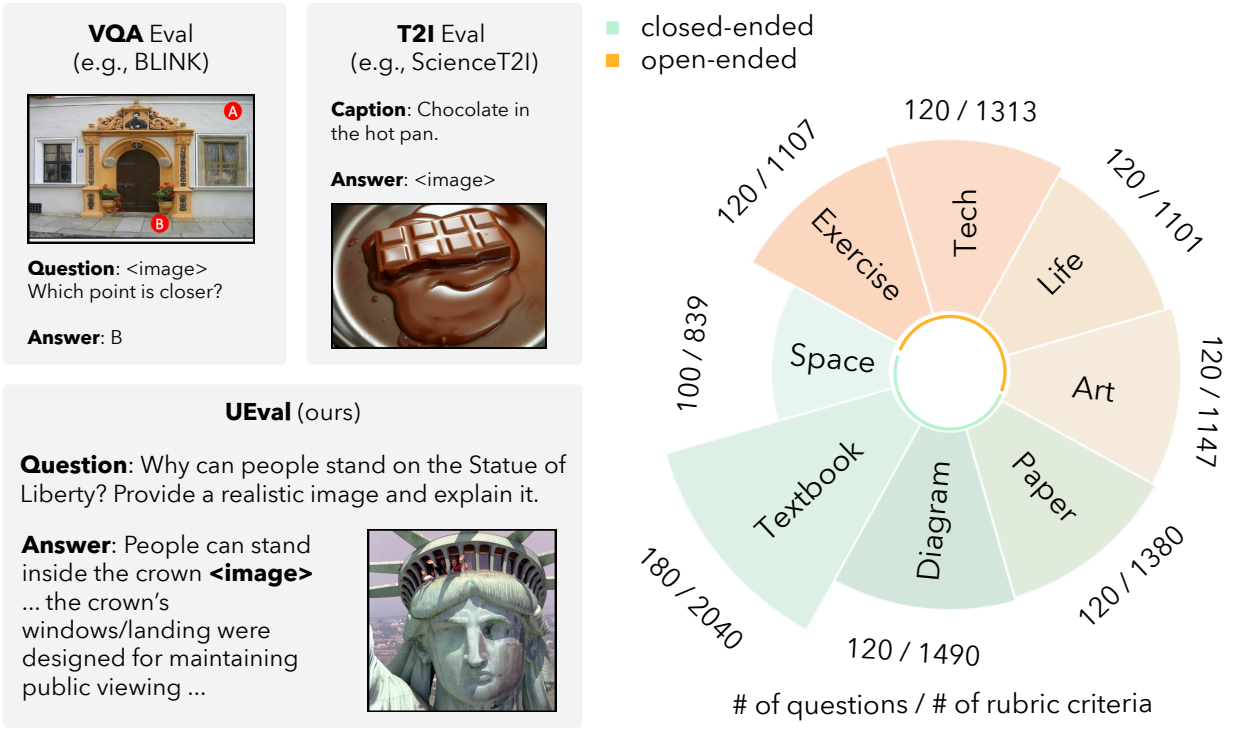

Figure 1 Left: Previous unified model evaluations focus on either image understanding (*i.e.*, VQA) or image generation from captions (*i.e.*, T2I). In contrast, UEval requires models to reason across modalities and generate responses in both images and text. The VQA example is from BLINK (Fu et al., 2024), and the T2I example is from ScienceT2I (Li et al., 2025a). **Right**: The chart illustrates the number of questions and rubric criteria across tasks in UEval.

UEval comprises 1,000 expert-curated questions spanning 8 diverse real-world tasks, including *space*, *textbook*, *paper*, *diagram*, *art*, *life*, *tech*, and *exercise*. Inspired by Arora et al. (2025), we propose a rubric-based framework for consistent and reproducible evaluation. For each question, we first manually collect reference answers in both text and image. Then, a frontier multimodal Large Language Model (MLLM) generates an initial rubric, consisting of multiple rubric criteria, conditioned on the original question and reference answers. Human annotators further refine these rubrics to eliminate redundancies and add any missing criteria. In total, UEval contains 10,417 rigorously validated rubric criteria to enable reliable automatic grading. In our experiments, we employ Gemini-2.5-Pro (Google, 2025a) as a judge model to score model responses with our rubrics and find that its scores show strong agreement with human judgments.

We conduct a comprehensive evaluation of 9 unified models on UEval. Our results show that UEval presents a challenge to all models. Among them, GPT-5-Thinking (OpenAI, 2025) achieves the highest score of 66.4 out of 100 when averaged across tasks, whereas the best-performing open-source model (*i.e.*, Emu3.5 (Cui et al., 2025)) reaches only 49.1. We also observe that current models struggle to generate multiple images with consistent labeling across steps in multi-step planning tasks (*e.g.*, drawing a cat step by step).

Interestingly, we observe that reasoning models (*e.g.*, GPT-5-Thinking) outperform their non-reasoning variants (*e.g.*, GPT-5-Instant) on most tasks. To further investigate the harm of reasoning in multimodal generation, we append the reasoning trace produced by GPT-5-Thinking to the end of the original question prompt and feed it into non-reasoning models. Surprisingly, this substantially improves the visual outputs generated by GPT-5-Instant and Gemini-2.5-Flash, while open-source models (*e.g.*, BAGEL) show no improvement. These observations suggest that Chain-of-Thought reasoning (Wei et al., 2022), long studied in Large Language Models (LLMs), may also play an important role in unified multimodal generation.

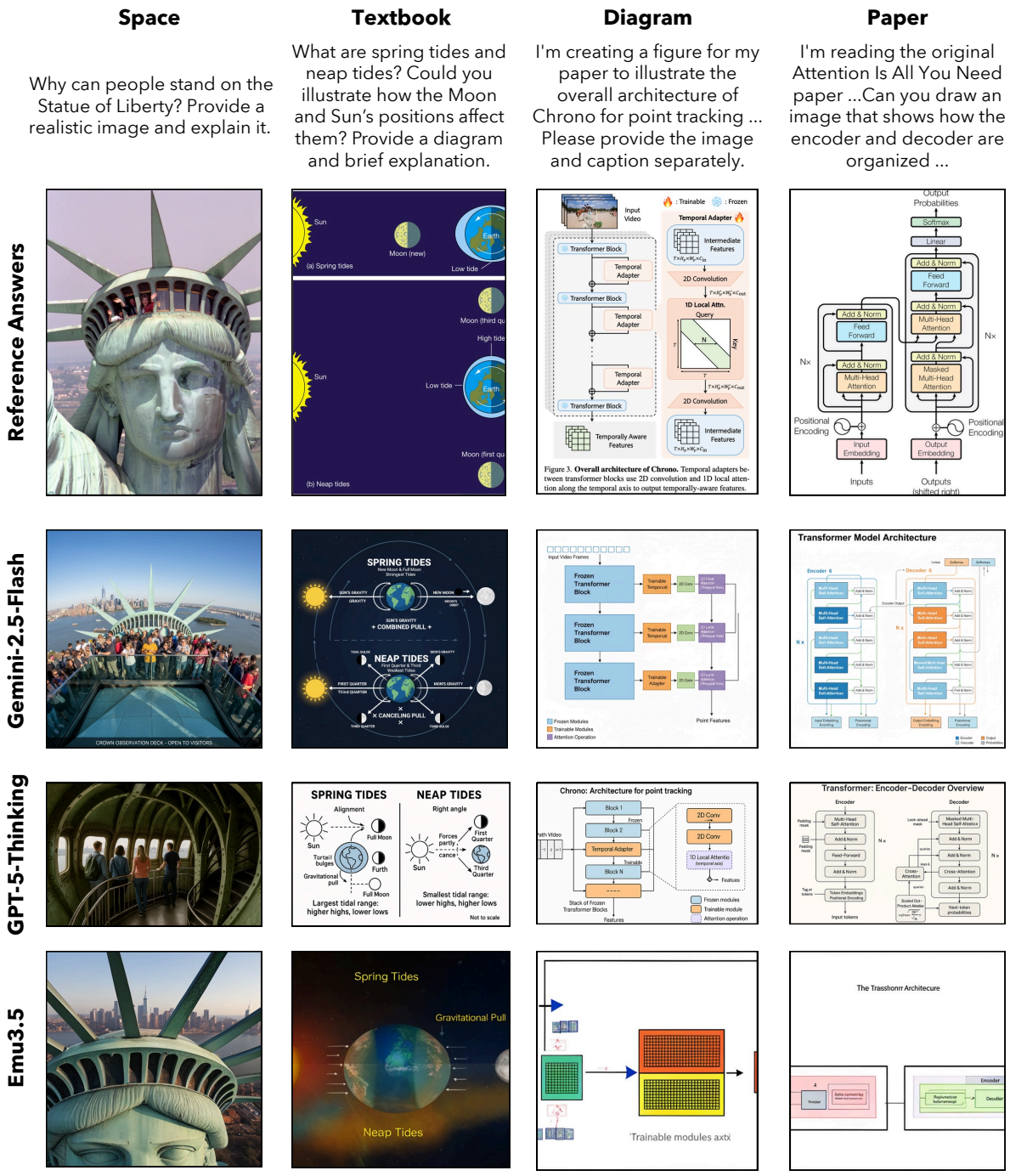

Figure 2 Model-generated images for closed-ended tasks in UEval. We visualize images generated by GPT-5-Thinking, Gemini-2.5-Flash, and Emu3.5 (Cui et al., 2025). These images fail to answer the questions accurately. For example, Gemini-2.5-Flash depicts a nonexistent external platform above the Statue of Liberty instead of the crown interior.

2 UEval

We present UEval, a benchmark designed to evaluate *unified multimodal* generation. UEval focuses on real-world tasks where a model must reason carefully before generating natural language and images in response to user queries. UEval consists of 8 tasks. These tasks fall into two groups: closed-ended tasks and open-ended tasks. They differ in both task design and evaluation dimensions used for model responses.

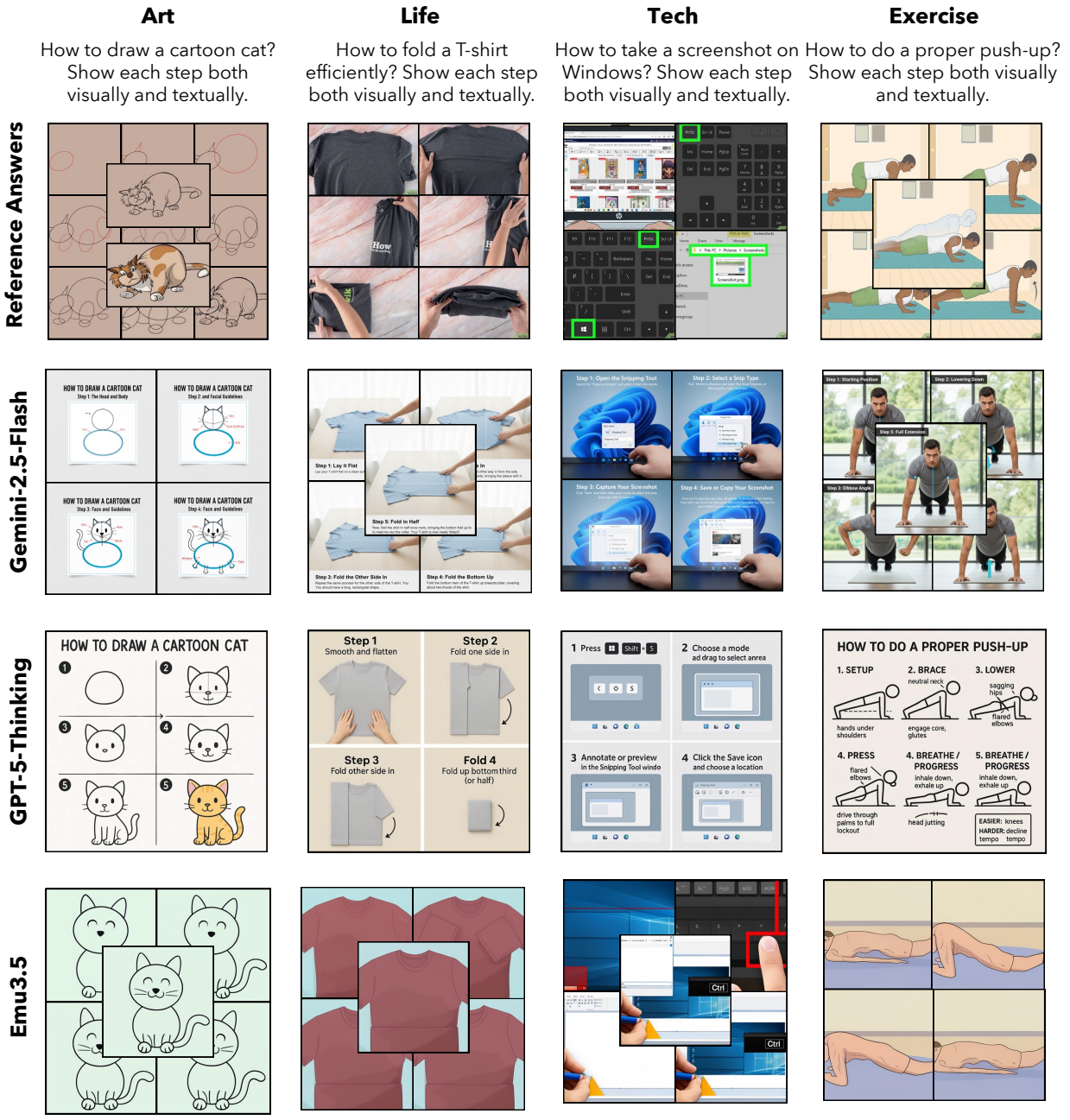

Figure 3 Model-generated images for open-ended tasks in UEval. We prompt GPT-5-Thinking, Gemini-2.5-Flash, and Emu3.5 to synthesize step-by-step visual guides for each task. The generated images often exhibit temporal inconsistencies. For instance, in the *art* task, when drawing a cartoon cat, GPT-5-Thinking mislabels sub-images (*e.g.*, two images tagged as step 5). For visualization, we stack the images generated by Gemini-2.5-Flash into a single grid.

Closed-ended tasks, including *space*, *textbook*, *diagram*, and *paper*, emphasize factual understanding and grounded explanation, and typically have a clear target answer that model outputs are expected to match. In contrast, open-ended tasks, including *art*, *life*, *tech*, and *exercise*, focus on step-by-step drawings that illustrate how to perform a task. Multiple plausible visualizations may exist for the same question in these tasks.

Figures 2 and 3 illustrate the 8 tasks in UEval alongside model-generated images. These tasks range from explanations with visual illustrations (*e.g.*, space) to academic figure creation (*e.g.*, paper). They also vary in format, from multi-step generation to single-step description (guide *vs.* diagram), and in breadth, from

general scientific knowledge to specialized academic content (textbook *vs.* paper).

Each sample in our benchmark includes a question prompt and a grading rubric with multiple rubric criteria to score model outputs. Our rubric criteria are drafted by Gemini-2.5-Pro (Google, 2025a) and then refined by humans (see Section 2.2 for more details). During evaluation, these rubrics are used by an MLLM judge to grade model responses for each question. In total, UEval contains 1,000 questions and 10,417 rubric criteria.

2.1 Dataset Composition

We describe each task in UEval. All open-ended tasks, including *art*, *life*, *tech*, and *exercise*, are grouped under a broader category, *guide*, as they share task design and data-collection procedure. Additional details on the image and text sources used to build each task are provided in Appendix B.

Space. This task evaluates a model's ability to depict specific architectural features. Generated images must highlight the structural elements relevant to the question rather than serve as decoration. For example, given the prompt *"Why can people stand on the Statue of Liberty? Provide a realistic image and explain it."*, a full-view image of the statue is insufficient, as it does not show the crown platform where visitors can stand.

To construct this task, we first collect a small set of 20 seed questions about well-known landmarks from online Q&A forums (*e.g.*, Quora-like platforms). Since there are not many such questions on the internet or in existing datasets, we use GPT-5 to expand this set. Given our seed questions, the model is prompted to propose additional questions for different landmarks. Human annotators then review all generated questions to ensure that each one refers to real, identifiable features of a landmark.

For every verified question, annotators retrieve a representative image from public sources (*e.g.*, Wikipedia) that depicts the relevant architectural feature and can answer the question visually. Finally, annotators write a short reference text describing how the chosen image illustrates the engineering features of that landmark.

Textbook. This task tests a model's ability to explain fundamental scientific phenomena (*e.g.*, geological transformations) through instructional diagrams. Generated outputs should identify underlying mechanisms (*e.g.*, DNA's role in genetics) or connections between concepts (*e.g.*, how a headland erodes into caves, arches, and stacks). An example question is *"I do not quite get how a headland turns into caves, arches, and stacks. Can you generate an image and explain the sequence? Please answer with both visual and textual explanations."*

We use the textbook-style diagrams and their answer texts in the TQA dataset (Kembhavi et al., 2017) as our reference image-text pairs. These pairs cover several subjects in science, including biology, geography, and chemistry. Building on these references, we prompt GPT-5 to generate learning-oriented questions that can be answered using our reference answers. Human annotators then manually review the generated questions to ensure that each is scientifically sound, unambiguous, and can be answered with the reference content.

Diagram. This task targets a common need in academic writing, where researchers design figures to illustrate complex methods in research papers. The evaluated model receives a technical description of a specific method or model architecture and is asked to synthesize a self-contained figure with a caption.

To construct this task, we manually curate figures from recent top-tier AI conference papers (*e.g.*, ICLR, CVPR) and pair each figure with its original caption as a reference image-text pair. We intentionally avoid diagrams from well-known papers to reduce the chance that evaluated models have seen them during training. We provide GPT-5 with the image as well as the full paper and prompt it to create a figure-generation instruction. This instruction captures important architectural or methodological details so that models can answer the question without needing access to the original paper. Human annotators then review and refine the instructions to check scientific correctness and relevance to the original figure.

Paper. This task assesses whether a unified model can accurately explain complex concepts from cutting-edge computer science research in an accessible way. Given a user question about a particular method, the model must first understand the technical content and then provide a clear, coherent explanation. An example question is *"I'm reading the original Attention Is All You Need paper and trying to understand the overall structure of the Transformer model. Can you draw an image that shows how the encoder and decoder are organized, and explain how data flows through the layers? Please answer with both visual and textual answers."*

We source a diverse set of figures from seminal papers (*e.g.*, Transformer (Vaswani et al., 2017), ResNet (He

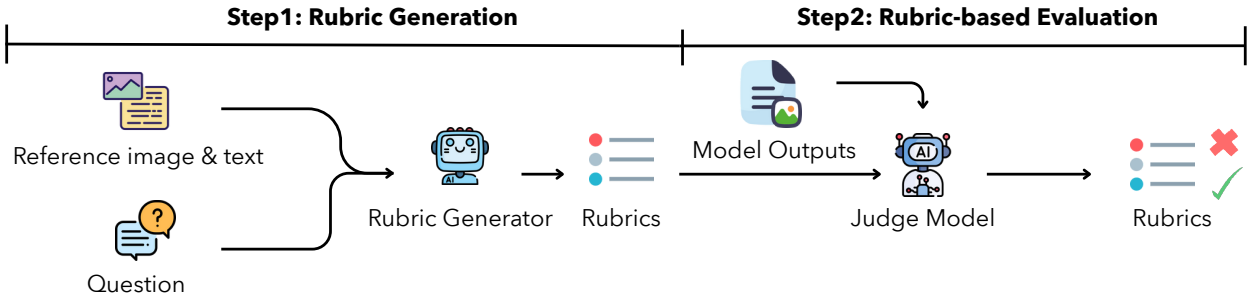

Figure 4 Rubric drafting and response evaluation procedure in UEval. We propose using *data-dependent* rubrics to evaluate outputs from unified models. For each question, a model drafts an itemized rubric based on the question and the reference image-text pair. A judge model then scores the generated response against each rubric criterion.

et al., 2016)) as well as online teaching platforms (*e.g.*, D2L (Zhang et al., 2023)). Each figure serves as a reference image. For each reference image, we extract the relevant technical descriptions from the original sources and prompt GPT-5 to generate reader-oriented questions together with explanatory reference text that answers those questions based on the figure. Human annotators then review the generated questions and answers to check technical validity and faithfulness to the original content.

Guide. This task evaluates a model's ability to produce a coherent, step-by-step visual guide for everyday activities. It contains four tasks, *art*, *life*, *tech*, and *exercise*, covering a range of real-world skills that require multi-step demonstration. For each question, the model must generate a visual guide (in one or more images) together with text explanations that illustrate a clear progression from an initial state to the final state. An example question from these tasks is *"How to draw a cartoon cat? Show each step both visually and textually."*

Questions and reference image-text answers are sourced from high-quality tutorial materials (*e.g.*, WikiHow) and step-by-step demonstrational videos (*e.g.*, YouTube). Although these tasks require visually illustrating specific skills through multi-step drawings, each reference answer consists of a sequence of images that progressively depict the drawing process and stepwise textual explanations. For the questions, we do not specify the number of steps and allow models to flexibly produce a sequence of images for each task.

2.2 Rubric Generation and Evaluation

How to effectively evaluate unified multimodal generation remains an open problem. Zhou et al. (2025b) use average win rates from pairwise comparisons of model outputs containing images and text, but this approach requires a large number of comparisons to obtain accurate model scores. These scores can change if a different set of models is evaluated. Moreover, most current evaluations (Xia et al., 2025; Zhou et al., 2025b) are *data-independent*: a single generic prompt is applied to grade all samples. As a result, this approach overlooks sample-specific differences and can lead to inaccurate results.

Rubric generation. Inspired by HealthBench (Arora et al., 2025), we adopt a *data-dependent* approach to evaluate unified multimodal generation. Figure 4 (*top*) illustrates our framework for generating rubrics and evaluating model outputs with an MLLM. For each sample, we provide the question along with its reference image-text answer to a rubric generator (*e.g.*, Gemini-2.5-Pro) to create a rubric with multiple fine-grained criteria that evaluate model outputs along different dimensions. For each question, about half of the rubric criteria are used to evaluate the generated image(s), while the others are used to evaluate the generated text.

Note that we prompt the rubric generator differently when drafting rubrics for closed-ended *vs.* open-ended tasks. Figure 5 illustrates this difference. For closed-ended tasks, rubrics check whether the generated content contains the key information present in the reference answers. In contrast, for open-ended tasks, rubrics evaluate higher-level qualities of the output (*e.g.*, temporal consistency). Rubrics for open-ended tasks include criteria such as *"The answer must describe a step-by-step process for drawing a cat"* and *"The sequence of images must illustrate the complete drawing process, from the initial basic shapes to the final version"*.

Human review. We conduct two rounds of human review to ensure the quality and reliability of all generated

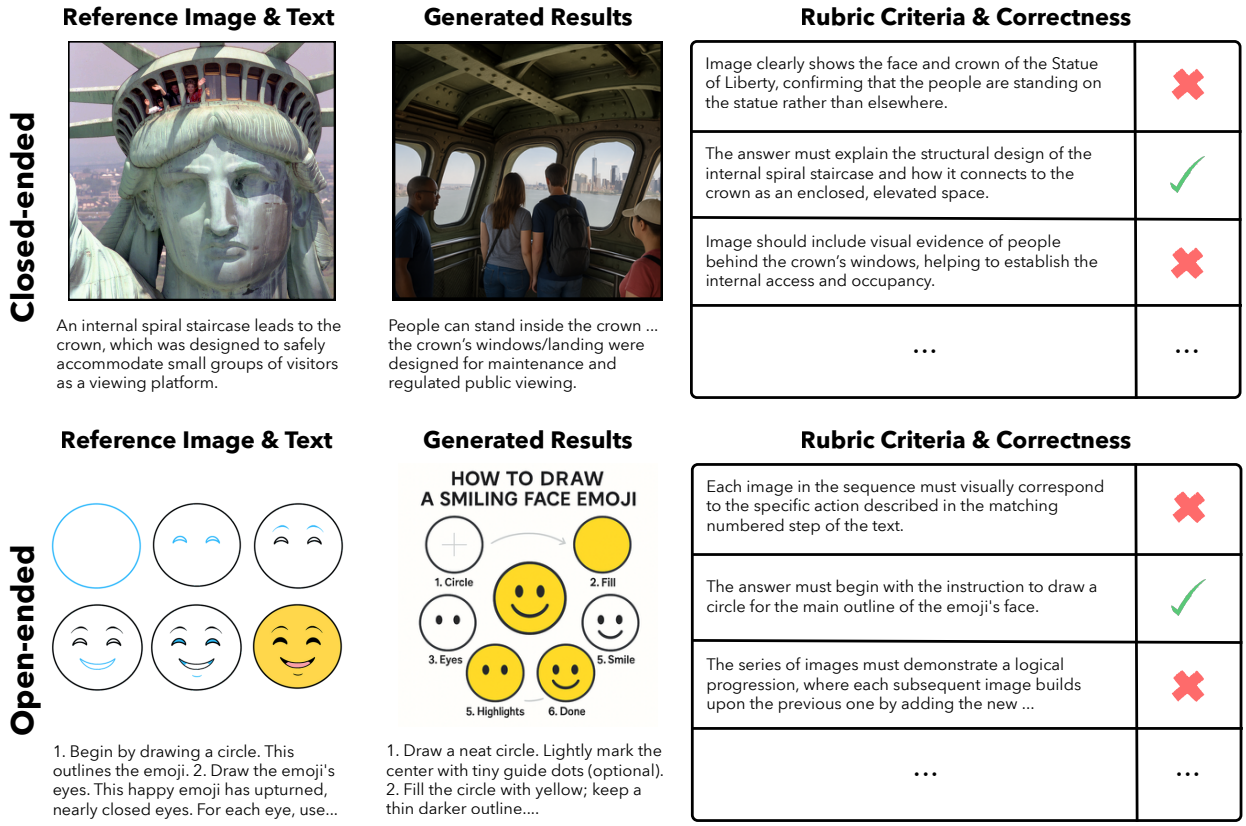

Figure 5 Rubric examples for closed-ended and open-ended tasks in UEval. Rubrics for closed-ended tasks check whether generated outputs contain specific details important for answering the questions, whereas rubrics for open-ended tasks evaluate higher-level generation qualities (*e.g.*, image-text alignment).

rubrics. Our goal is to design rubrics that correctly reward responses similar to the reference answers while penalizing erroneous ones. For each benchmark sample, a primary annotator first verifies the question, reference answer, and rubrics for correctness and alignment with the task. Subsequently, other co-authors independently review the annotations. Only rubric items unanimously judged by all reviewers to be unambiguous and well aligned with the task design are retained. Overall, human annotators supervise benchmark construction at a system level, from the inputs (questions) to the outputs (reference answers) and the grading criteria (rubrics).

Human annotators refine the model-generated rubric items in several ways. First, repeated or similar rubric criteria are consolidated. For example, we merge *"the steps must be sequential"* and *"the steps should follow a logical order"* into a single, more precise requirement. Second, we add important but missing rubric criteria. For example, in questions involving rendered text, we introduce additional rubric criteria to evaluate overall text generation quality, such as *"all visible text in the generated image(s) must be spelled correctly and rendered naturally, with no misspellings, garbled characters, distortions, or nonsensical text"*.

Rubric-based evaluation. We employ an MLLM (*e.g.*, Gemini-2.5-Pro) as a judge to evaluate images and text generated by unified models. For each sample, the judge checks the model response against each rubric criterion, and the final score is computed as the fraction of satisfied rubric criteria over the total number of rubric criteria. This provides an automated, reproducible evaluation method in place of human judgment. We find that using a frontier model yields results well aligned with human judgment, and that some strong judge models produce similar scores. We will discuss this further in Section 3.3.

In Table 1 (*gray*), we report the scores of the reference answers graded with our rubrics, using Gemini-2.5-Pro as the judge model. Overall, the reference image-text answers achieve an average rubric score of 92.2 across all tasks, indicating that our rubrics capture most of the important features of the reference answers. Nevertheless, we observe that reference answers for open-ended tasks receive slightly lower scores than those for closed-ended

	Space	Textbook	Diagram	Paper	Art	Life	Tech	Exercise	Avg
Reference	96.2	94.4	93.1	96.2	90.6	87.7	90.6	89.2	92.2
Open-source Models									
Janus-Pro	21.0	31.0	37.4	15.2	26.4	23.0	17.6	11.5	22.9
Show-o2	25.4	33.1	33.2	17.4	25.6	15.6	17.4	13.1	22.6
MMaDA	10.8	20.0	14.2	13.3	15.7	15.8	12.4	12.6	14.4
BAGEL	29.8	42.5	37.2	20.0	39.0	33.6	24.8	21.4	31.0
Emu3.5	**59.1**	**57.4**	**41.1**	**31.6**	**59.3**	**62.0**	**37.0**	**45.4**	**49.1**
Proprietary Frontier Models									
Gemini-2.0-Flash	65.2	55.2	47.6	45.8	**70.4**	58.0	50.2	48.0	55.1
Gemini-2.5-Flash	78.0	74.0	66.4	**71.6**	66.6	63.0	**58.2**	50.0	66.0
GPT-5-Instant	77.3	77.9	62.3	55.1	71.2	**69.7**	50.7	57.6	65.2
GPT-5-Thinking	**84.0**	**78.0**	**67.8**	51.9	67.8	63.8	57.0	**61.4**	**66.4**

Table 1 UEval leaderboard. We evaluate open-source and proprietary frontier models on 8 tasks in UEval. **Bold** indicates the highest performance for each column within each group (*e.g.*, open-source *vs.* proprietary frontier).

tasks. This is likely because reference answers for open-ended tasks contain multiple images, making them harder for the judge model to evaluate accurately, whereas closed-ended tasks involve only a single image. We expect that more capable MLLMs will further improve the effectiveness of our rubric-based evaluation.

3 Experiments

3.1 Settings

Models. We evaluate recent unified models on all 8 tasks in our benchmark. For open-source models, we consider Janus-Pro (Chen et al., 2025d), Show-o2 (Xie et al., 2025), MMaDA (Yang et al., 2025), BAGEL (Deng et al., 2025), and Emu3.5 (Cui et al., 2025). For proprietary frontier models, we evaluate Gemini-2.0-Flash (Google, 2024), Gemini-2.5-Flash (*i.e.*, Nano Banana) (Google, 2025b), GPT-5-Instant (OpenAI, 2025), and GPT-5-Thinking (OpenAI, 2025). We access both GPT-5 models through the official chat interface.

Evaluation setup. Some models (*e.g.*, Janus-Pro, Show-o2, MMaDA, and BAGEL) can generate only images or text by design, but not both in a single inference pass. To obtain both outputs, we feed the same question prompt to the model twice, once to generate the image and once to generate the text. For models that natively support joint image-text generation (*e.g.*, GPT-5, Gemini), we directly collect their multimodal responses. We use Gemini-2.5-Pro (Google, 2025a) to grade the generated outputs based on fine-grained rubrics (Section 2.2). Appendix C provides the full prompt to evaluate model responses.

3.2 Results

Table 1 reports the performance of various models on UEval. Overall, frontier models consistently outperform open-source ones across all tasks: GPT-5-Thinking achieves the highest average score of 66.4, while the best open-source model obtains only 49.1. To better understand performance differences, Figure 6 presents a radar chart comparing models across tasks. The gap between proprietary and open-source models is large: the strongest frontier model (*i.e.*, GPT-5-Thinking) outperforms the best open-source model (*i.e.*, Emu3.5) by over 17 points on average. The individual image and text scores for each task are provided in Appendix A.

We also observe that the tasks requiring multi-step planning (*e.g.*, art, life) yield substantially lower scores than knowledge-based tasks (*e.g.*, textbook, diagram). Figure 3 further illustrates this pattern. For example, in the *art* task, GPT-5-Thinking incorrectly labels the final two images as *step 5*. Similar mistakes also occur in both the *life* and *exercise* tasks. Likewise, Gemini-2.5-Flash changes the shirt's orientation from step 1 to step 2 and then changes it back in step 3 in the *life* task. More interestingly, current *reasoning* models (*e.g.*, GPT-5-Thinking) achieve much better performance than non-reasoning ones (*e.g.*, GPT-5-Instant). We further study the benefits of *reasoning* in multimodal generation in the next section.

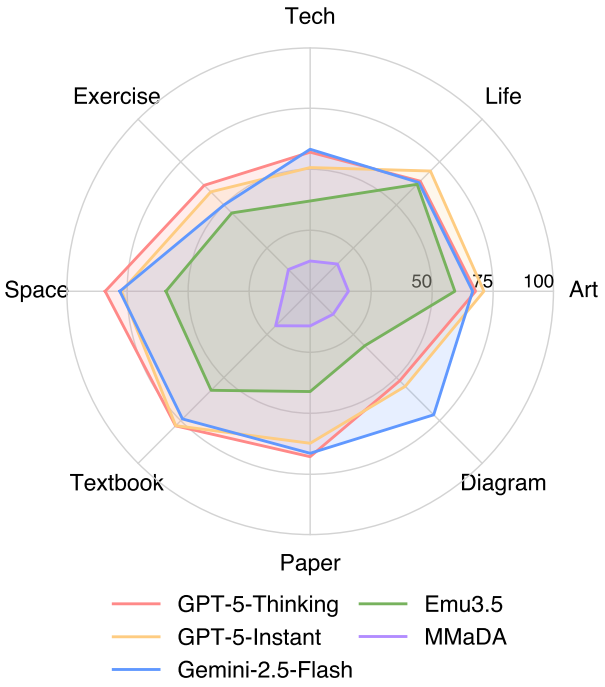

Figure 6 Model performance on UEval. Each axis corresponds to a task, and each colored polygon is a different model. Proprietary frontier models (*e.g.*, GPT-5-Thinking) consistently outperform open-source models (*e.g.*, Emu3.5).

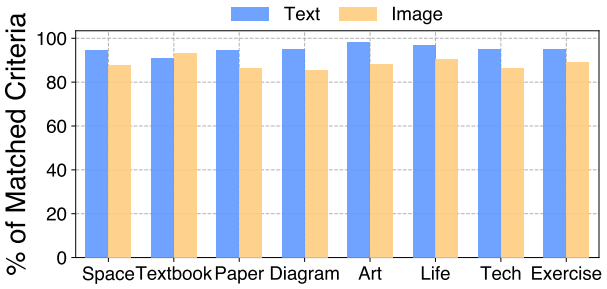

Figure 7 Percentage of matched rubric criteria between human evaluators and an MLLM judge. We find high agreement between an MLLM and humans.

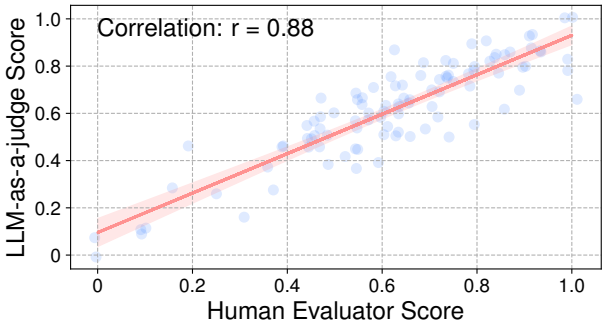

Figure 8 LLM-as-a-judge *vs.* human evaluator. Each point represents a score for a question in UEval. The Pearson correlation (r = 0.88) demonstrates that the LLM judge closely aligns with human evaluation.

3.3 Analysis

Performance on text and image generation. As described in Section 2.2, each question contains multiple rubric criteria, with about half evaluating generated images and the others evaluating generated text. Table 2 reports separate image and text scores for all evaluated models (from Table 1), averaged across all questions to assess text and image generation performance separately. We also report text scores for text-only models. Overall, current models show stronger performance in text generation than in image generation. Moreover, for open-source unified m odels, t heir t ext generation ability still lags behind that of text-only models, highlighting that achieving high-quality multimodal generation is still a significant challenge and that substantial room for worsening remains.

	Image	Text
Reference	87.8	96.8
Open-source Models		
Janus-Pro	6.5	39.3
Show-o2	8.3	36.9
MMaDA	15.9	12.8
BAGEL	13.6	48.5
Emu3.5	**33.6**	**64.6**
Proprietary Frontier Models		
Gemini-2.0-Flash	36.9	73.2
Gemini-2.5-Flash	**56.4**	75.5
GPT-5-Instant	52.8	77.7
GPT-5-Thinking	49.1	**83.8**
Text-only Models		
Qwen3-32B	–	81.2
Gemma3-27B	–	83.1

Table 2 Image and text scores on UEval, averaged over all 8 tasks. We find a performance gap between image and text generation. We also report the text scores of text-only models for comparison.

Human evaluation. To assess the reliability of using an MLLM as a judge, we randomly sample 10% of GPT-5-Thinking outputs from each task and ask human annotators to determine how many rubric criteria each model response satisfies. In Figure 7, we report the percentage of rubric criteria on which human evaluators and the LLM-as-a-judge agree. We observe approximately 90% of rubric criteria are matched between human evaluators and the LLM-as-a-judge across tasks. LLM-as-a-judge grading aligns very closely with human judgment across different tasks. Moreover, in Figure 8, we plot LLM-as-a-judge scores against human evaluator scores and observe a high Pearson correlation (r = 0.88). These results indicate that our automated evaluation framework is robust aligned with human judgment.

Different judge models. Our default judge model in Table 1 is Gemini-2.5-Pro (Google, 2025a). To understand how different judge models affect scores, we grade model responses generated by GPT-5-Thinking using other proprietary frontier and open-source models, including GPT-5-Thinking (OpenAI, 2025), Seed1.6-Vision (Seed, 2025), Qwen3-VL-235B-Thinking/Instruct (Bai et al., 2025), and GLM-4.1V-Thinking (GLM, 2025).

	Space	Textbook	Diagram	Paper	Art	Life	Tech	Exercise
Gemini-2.5-Pro	84.0	78.0	67.8	51.9	67.8	63.8	57.0	61.4
GPT-5-Thinking	80.0	73.0	55.8	46.3	56.4	50.1	45.4	51.7
Seed1.6-Vision	85.5	81.0	68.2	53.8	75.2	70.8	67.8	70.9
Qwen3-VL-235B-Thinking	81.8	78.4	63.4	49.3	56.8	56.8	53.6	59.6
Qwen3-VL-235B-Instruct	82.0	85.2	72.3	53.6	73.7	61.4	57.7	63.0
GLM-4.1V-Thinking	84.3	83.6	68.2	49.8	79.4	74.5	74.3	70.7

Table 3 Scores of GPT-5-Thinking responses evaluated by different judge models. Gemini-2.5-Pro, GPT-5-Thinking, and Qwen3-VL-235B-Thinking produce consistent scores across tasks, whereas other models yield very different scores.

Table 3 reports the per-task scores on GPT-5-Thinking responses graded with different judge models. We observe that GPT-5-Thinking, Gemini-2.5-Pro, and Qwen3-VL-235B-Thinking produce similar scores across all tasks, whereas other models (*e.g.*, Seed1.6-Vision, GLM-4.1V-Thinking) yield very different ones, especially in open-ended tasks (*e.g.*, *art*, *life*). Therefore, we recommend using GPT-5-Thinking, Gemini-2.5-Pro, or Qwen3-VL-235B-Thinking as the judge model for grading model-generated responses in UEval. **The effectiveness of reasoning traces.** To understand why reasoning models (*e.g.*, GPT-5-Thinking) yield better results in multimodal generation, we record a reasoning trace and append it to the original question prompt. We then provide non-reasoning models (*e.g.*, GPT-5-Instant) with this modified prompt. Figure 9 visualizes images generated by some non-reasoning models. Surprisingly, incorporating the reasoning trace enables GPT-

Figure 9 Reasoning can improve the quality of multimodal generation. We extract the *reasoning trace* from GPT-5-Thinking and append it to the original question. The resulting prompt is then provided to *non-reasoning* models (*e.g.*, GPT-5-Instant, Gemini-2.5-Flash, and BAGEL) to generate responses. This can steer the generated images toward that of the reasoning model and lead to more accurate responses, with BAGEL as an exception.

5-Instant and Gemini-2.5-Flash to generate a more accurate image of the interior of the Statue of Liberty's crown. This suggests that multimodal generation in unified models benefits from Chain-of-Thought (Wei et al., 2022) reasoning generated by other models. However, this does not apply to all unified models: weaker models (*e.g.*, BAGEL) do not benefit from this added reasoning. A sufficiently strong multimodal generation capability is necessary to effectively leverage such additional reasoning signals.

4 Related Work

Multimodal large language models and unified models. Multimodal Large Language Models (MLLMs) have progressed greatly in recent years. These models (Alayrac et al., 2022; Dai et al., 2023; Liu et al., 2023; Li et al., 2023; Zhu et al., 2024) typically integrate a visual encoder (Radford et al., 2021; Dosovitskiy et al., 2021; He et al., 2022) with a pre-trained Large Language Model (Llama-2-Team, 2023; Peng et al., 2023), achieving strong performance on image captioning (Chen et al., 2015; Plummer et al., 2015) and visual question answering (Goyal et al., 2017; Mathew et al., 2021). To further scale their capabilities, these models require millions of instruction-tuning data (Tong et al., 2024a; Deitke et al., 2025).

A parallel line of research seeks to unify multimodal understanding and generation within a single model. Some methods adopt diffusion-based approaches (Dong et al., 2024; Yang et al., 2025; Li et al., 2025c; Shi et al., 2025a; Wang et al., 2025), whereas others train models with a purely autoregressive objective (Team, 2024; Wang et al., 2024; Wu et al., 2025a,b; Qu et al., 2025). There are also hybrid methods that combine both approaches (Deng et al., 2025; Zhou et al., 2025a; Xie et al., 2025; Chen et al., 2025b). We refer readers to Zhang et al. (2025) for a comprehensive survey of MLLMs and unified models.

Multimodal benchmarks. A range of benchmarks has been proposed to evaluate multimodal inputs. Initial work (Goyal et al., 2017; Marino et al., 2019; Masry et al., 2022) evaluates image understanding for specific image types, and later efforts benchmark broader image coverage (Liu et al., 2024b; Yue et al., 2024). There are also studies evaluating text-to-image generation quality (Saharia et al., 2022; Huang et al., 2023; Ghosh et al., 2023; Lin et al., 2024). Some benchmarks (*e.g.*, VDC (Chai et al., 2025)) begin to use data-dependent rubrics for better evaluation. More recent works unify understanding and generation benchmarks as interleaved text-and-image generation (An et al., 2024; Liu et al., 2024a; Chen et al., 2025a) or unified multimodal generation (Li et al., 2025b; Zou et al., 2025; Shi et al., 2025b). In contrast to them, our evaluation is very simple: an MLLM judge is used to grade model responses based on rubrics. This avoids per-sample human scoring (Zhou et al., 2025b) or training a scoring model (Xia et al., 2025) for evaluation.

5 Conclusion

We introduce UEval, a benchmark to evaluate unified multimodal generation beyond standard tasks (*e.g.*, visual question answering, text-to-image generation). Our benchmark contains 1,000 samples across 8 real-world tasks and provides 10,417 fine-grained rubric criteria for rigorous, automated grading of model responses. Our results demonstrate that UEval is challenging for both proprietary frontier and open-source unified models. We also observe that reasoning can improve multimodal generation quality. We hope this work will stimulate further research on developing stronger models and better benchmarks for multimodal generation.

6 Acknowledgments

We gratefully acknowledge the use of the Neuronic GPU computing cluster maintained by the Department of Computer Science at Princeton University. This work was performed using Princeton Research Computing resources, a consortium led by the Princeton Institute for Computational Science and Engineering (PICSciE) and Research Computing at Princeton University.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022.

Jie An, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Lijuan Wang, and Jiebo Luo. Openleaf: A novel benchmark for open-domain interleaved image-text generation. In *ACM MM*, 2024.

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher D Manning. Auroracap: Efficient, performant video detailed captioning and a new benchmark. In *ICLR*, 2025.

Dongping Chen, Ruoxi Chen, Shu Pu, Zhaoyi Liu, Yanru Wu, Caixi Chen, Benlin Liu, Yue Huang, Yao Wan, Pan Zhou, et al. Interleaved scene graphs for interleaved text-and-image generation assessment. In *ICLR*, 2025a.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025b.

Liang Chen, Shuai Bai, Wenhao Chai, Weichu Xie, Haozhe Zhao, Leon Vinci, Junyang Lin, and Baobao Chang. Multimodal representation alignment for image generation: Text-image interleaved control is easier than you think. *arXiv preprint arXiv:2502.20172*, 2025c.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025d.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, Yueze Wang, Chengyuan Wang, Fan Zhang, Yingli Zhao, Ting Pan, Xianduo Li, Zecheng Hao, Wenxuan Ma, Zhuo Chen, Yulong Ao, Tiejun Huang, Zhongyuan Wang, and Xinlong Wang. Emu3.5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583*, 2025.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *CVPR*, 2025.

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.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, Xiangyu Zhang, Kaisheng Ma, and Li Yi. Dreamllm: Synergistic multimodal comprehension and creation. In *ICLR*, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *NeurIPS*, 2025.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *ECCV*, 2024.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *NeurIPS*, 2023.

GLM. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Google. Introducing gemini 2.0: our new ai model for the agentic era. *https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/*, 2024.

Google. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv: 2507.06261*, 2025a.

Google. Introducing gemini 2.5 flash image, our state-of-the-art image model. *https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/*, 2025b.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In *NeurIPS*, 2023.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *CVPR*, 2017.

Jialuo Li, Wenhao Chai, Xingyu Fu, Haiyang Xu, and Saining Xie. Science-t2i: Addressing scientific illusions in image synthesis. In *CVPR*, 2025a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.

Yi Li, Haonan Wang, Qixiang Zhang, Boyu Xiao, Chenchang Hu, Hualiang Wang, and Xiaomeng Li. Unieval: Unified holistic evaluation for unified multimodal understanding and generation. *arXiv preprint arXiv:2505.10483*, 2025b.

Zijie Li, Henry Li, Yichun Shi, Amir Barati Farimani, Yuval Kluger, Linjie Yang, and Peng Wang. Dual diffusion for unified image generation and understanding. In *CVPR*, 2025c.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *ECCV*, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

Minqian Liu, Zhiyang Xu, Zihao Lin, Trevor Ashby, Joy Rimchala, Jiaxin Zhang, and Lifu Huang. Holistic evaluation for interleaved text-and-image generation. In *EMNLP*, 2024a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *ECCV*, 2024b.

Llama-2-Team. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, 2019.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *ACL*, 2022.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, 2021.

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.

OpenAI. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/, 2025.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015.

Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. In *CVPR*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022.

Seed. Seed1.6. https://seed.bytedance.com/en/seed1_6, 2025.

Qingyu Shi, Jinbin Bai, Zhuoran Zhao, Wenhao Chai, Kaidong Yu, Jianzong Wu, Shuangyong Song, Yunhai Tong, Xiangtai Li, Xuelong Li, et al. Muddit: Liberating generation beyond text-to-image with a unified discrete diffusion model. *arXiv preprint arXiv:2505.23606*, 2025a.

Yang Shi, Yuhao Dong, Yue Ding, Yuran Wang, Xuanyu Zhu, Sheng Zhou, Wenting Liu, Haochen Tian, Rundong Wang, Huanqian Wang, Zuyan Liu, Bohan Zeng, Ruizhe Chen, Qixun Wang, Zhuoran Zhang, Xinlong Chen, Chengzhuo Tong, Bozhou Li, Chaoyou Fu, Qiang Liu, Haotian Wang, Wenjing Yang, Yuanxing Zhang, Pengfei Wan, Yi-Fan Zhang, and Ziwei Liu. Realunify: Do unified models truly benefit from unification? a comprehensive benchmark. *arXiv preprint arXiv:2509.24897*, 2025b.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *NeurIPS*, 2024a.

Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. In *ICCV*, 2024b.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

Jin Wang, Yao Lai, Aoxue Li, Shifeng Zhang, Jiacheng Sun, Ning Kang, Chengyue Wu, Zhenguo Li, and Ping Luo. Fudoki: Discrete flow-based unified understanding and generation via kinetic-optimal velocities. *arXiv preprint arXiv:2505.20147*, 2025.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *CVPR*, 2025a.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, Song Han, and Yao Lu. Vila-u: a unified foundation model integrating visual understanding and generation. In *ICLR*, 2025b.

Peng Xia, Siwei Han, Shi Qiu, Yiyang Zhou, Zhaoyang Wang, Wenhao Zheng, Zhaorun Chen, Chenhang Cui, Mingyu Ding, Linjie Li, et al. Mmie: Massive multimodal interleaved comprehension benchmark for large vision-language models. In *ICLR*, 2025.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. In *NeurIPS*, 2025.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. In *NeurIPS*, 2025.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024.

Aston Zhang, Zachary C Lipton, Mu Li, and Alexander J Smola. Dive into deep learning. *https://d2l.ai*, 2023.

Xinjie Zhang, Jintao Guo, Shanshan Zhao, Minghao Fu, Lunhao Duan, Jiakui Hu, Yong Xien Chng, Guo-Hua Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, and Kaifu Zhang. Unified multimodal understanding and generation models: Advances, challenges, and opportunities. *arXiv preprint arXiv:2505.02567*, 2025.

Xiangyu Zhao, Peiyuan Zhang, Kexian Tang, Xiaorong Zhu, Hao Li, Wenhao Chai, Zicheng Zhang, Renqiu Xia, Guangtao Zhai, Junchi Yan, et al. Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing. In *NeurIPS*, 2025.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *ICLR*, 2025a.

Pengfei Zhou, Xiaopeng Peng, Jiajun Song, Chuanhao Li, Zhaopan Xu, Yue Yang, Ziyao Guo, Hao Zhang, Yuqi Lin, Yefei He, et al. Opening: A comprehensive benchmark for judging open-ended interleaved image-text generation. In *CVPR*, 2025b.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *ICLR*, 2024.

Kai Zou, Ziqi Huang, Yuhao Dong, Shulin Tian, Dian Zheng, Hongbo Liu, Jingwen He, Bin Liu, Yu Qiao, and Ziwei Liu. Uni-mmmu: A massive multi-discipline multimodal unified benchmark. *arXiv preprint arXiv:2510.13759*, 2025.

Appendix

A Individual Image and Text Scores

As detailed in Section 2.2, we construct two rubrics for every question in UEval, one evaluating images and one evaluating text. Table 4 reports the corresponding per-modality scores for the results in Table 1. Across all models, we observe that text scores are much higher than image scores. This is especially pronounced for open-source models, which generate reasonably good text but perform extremely poorly on image generation, often receiving <10 image scores. In contrast, frontier models show a much narrower gap between their image and text performance, indicating more balanced multimodal generation capabilities.

	Space		**Textbook**		**Diagram**		**Paper**		**Avg**	
	Image	Text	Image	Text	Image	Text	Image	Text	Image	Text
Reference	94.7	97.8	92.5	96.2	92.2	93.9	94.7	97.8	93.5	96.4
Open-source Models										
Janus-Pro	13.5	28.5	10.8	51.3	0.8	73.9	3.3	27.2	7.1	45.2
Show-o2	22.7	28.1	10.9	55.2	9.3	57.2	4.6	30.2	11.9	42.7
MMaDA	5.0	16.7	17.7	22.4	9.2	19.2	**15.6**	11.0	11.9	17.3
BAGEL	27.6	31.9	14.5	70.6	2.8	71.7	4.8	35.2	12.4	52.3
Emu3.5	**64.6**	**53.6**	**32.5**	**82.3**	**13.3**	**68.9**	12.3	**50.8**	**30.7**	**63.9**
Proprietary Frontier Models										
Gemini-2.0-Flash	59.9	70.5	29.9	80.5	21.9	73.3	17.3	74.4	32.2	74.7
Gemini-2.5-Flash	**82.6**	73.5	61.5	86.6	**55.7**	77.2	**59.7**	**83.4**	**64.9**	**80.2**
GPT-5-Instant	74.9	79.7	**67.4**	88.5	44.5	80.1	27.4	82.8	53.6	82.8
GPT-5-Thinking	82.3	**85.7**	65.9	**90.0**	51.2	**84.3**	43.3	60.5	60.7	80.1

(a) closed-ended tasks

	Art		**Life**		**Tech**		**Exercise**		**Avg**	
	Image	Text	Image	Text	Image	Text	Image	Text	Image	Text
Reference	82.5	98.8	78.3	97.1	87.0	94.2	80.1	98.2	82.0	97.1
Open-source Models										
Janus-Pro	8.9	44.0	5.9	40.0	4.4	30.8	4.1	18.9	5.8	33.4
Show-o2	5.8	45.4	4.2	26.9	4.6	30.2	4.0	22.2	4.6	31.2
MMaDA	18.9	12.5	21.5	10.1	**17.3**	7.5	22.1	3.0	20.0	8.3
BAGEL	19.8	58.2	19.1	48.1	5.0	44.6	15.2	27.7	14.8	44.6
Emu3.5	**39.0**	**79.6**	**53.6**	**70.4**	16.2	**57.8**	**37.6**	**53.3**	**36.6**	**65.3**
Proprietary Frontier Models										
Gemini-2.0-Flash	55.5	85.2	44.9	71.2	30.1	70.4	35.7	60.4	41.6	71.8
Gemini-2.5-Flash	51.2	81.9	55.2	70.8	**48.9**	67.4	36.6	63.4	48.0	70.9
GPT-5-Instant	**58.9**	83.4	**60.7**	78.8	34.3	67.1	**54.2**	61.0	**52.0**	72.6
GPT-5-Thinking	40.8	**94.7**	41.8	**85.8**	26.2	**87.9**	41.2	**81.5**	37.5	**87.5**

(b) open-ended tasks

Table 4 Image and text scores on 8 tasks in UEval. Each task is evaluated separately for image and text generation. The results show that text generation consistently outperforms image generation.

B Data sources

Table 5 lists the data sources of the questions, reference images, and reference texts for each task in UEval.

Task	Question Source	Reference Image Source	Reference Text Source
Space	Quora, GPT-5 generated	Wikipedia, TripAdvisor, Google search	Wikipedia, TripAdvisor, Quora
Textbook	GPT-5 generated	TQA dataset	TQA dataset
Diagram	GPT-5 generated	arXiv, ICLR, CVPR, NeurIPS, ICCV	arXiv, ICLR, CVPR, NeurIPS, ICCV
Paper	GPT-5 generated	arXiv, D2L, Medium, CS231n	arXiv, D2L, Medium, CS231n
Art	WikiHow, EasyDrawing, YouTube, ArtForKidsHub	WikiHow, EasyDrawing, YouTube, ArtForKidsHub	WikiHow, EasyDrawing, YouTube, ArtForKidsHub
Life	WikiHow, YouTube	WikiHow, YouTube	WikiHow, YouTube
Tech	WikiHow, Food52, YouTube	WikiHow, Food52, YouTube	WikiHow, Food52, YouTube
Exercise	WikiHow, Healthline, Men's Health, YouTube	WikiHow, Healthline, Men's Health, YouTube	WikiHow, Healthline, Men's Health, YouTube

Table 5 Per-task breakdown of question, reference image, and reference text sources. The table shows the diverse origins of data used for different tasks in UEval.

C Evaluation Details

Figure 10 shows the prompt used to grade model responses (*i.e.*, image + text) with an MLLM judge (*e.g.*, Gemini-2.5-Pro) based on a specific rubric item during evaluation.

You are an expert AI evaluator. Your job is to look at a conversation, a rubric item and assess model outputs (text and images) against specific rubric criteria.Return a JSON object with "criteria_met".

Conversation
Question: <<question>>
Text Answer: <<text_answer>>
Image Answer: <<image_answer>>

Rubric Item
<<rubric_item>>

Instructions
Return a JSON object with "criteria_met".
- Set "criteria_met" to true only if the rubric is completely satisfied; false otherwise. Use "not sure" if the evidence is insufficient.
- One important clarification regarding the requirement is that each image must include a visual depiction of the described action – it cannot rely solely on text rendered within the image as a substitute for visual content.
- One important exception to the above point is that when the criterion is used to evaluate the consistency between an image step and its corresponding text step, the image does not need to depict all actions or details mentioned in that step to meet the criterion.

Figure 10 Evaluation prompt used for Gemini-2.5-Pro as the judge model. It shows how a single rubric item is applied to evaluate model responses and produce structured judgments.

D Responses from More Models

Figures 11, 12, 13, and 14 visualize additional examples of images generated by different models.

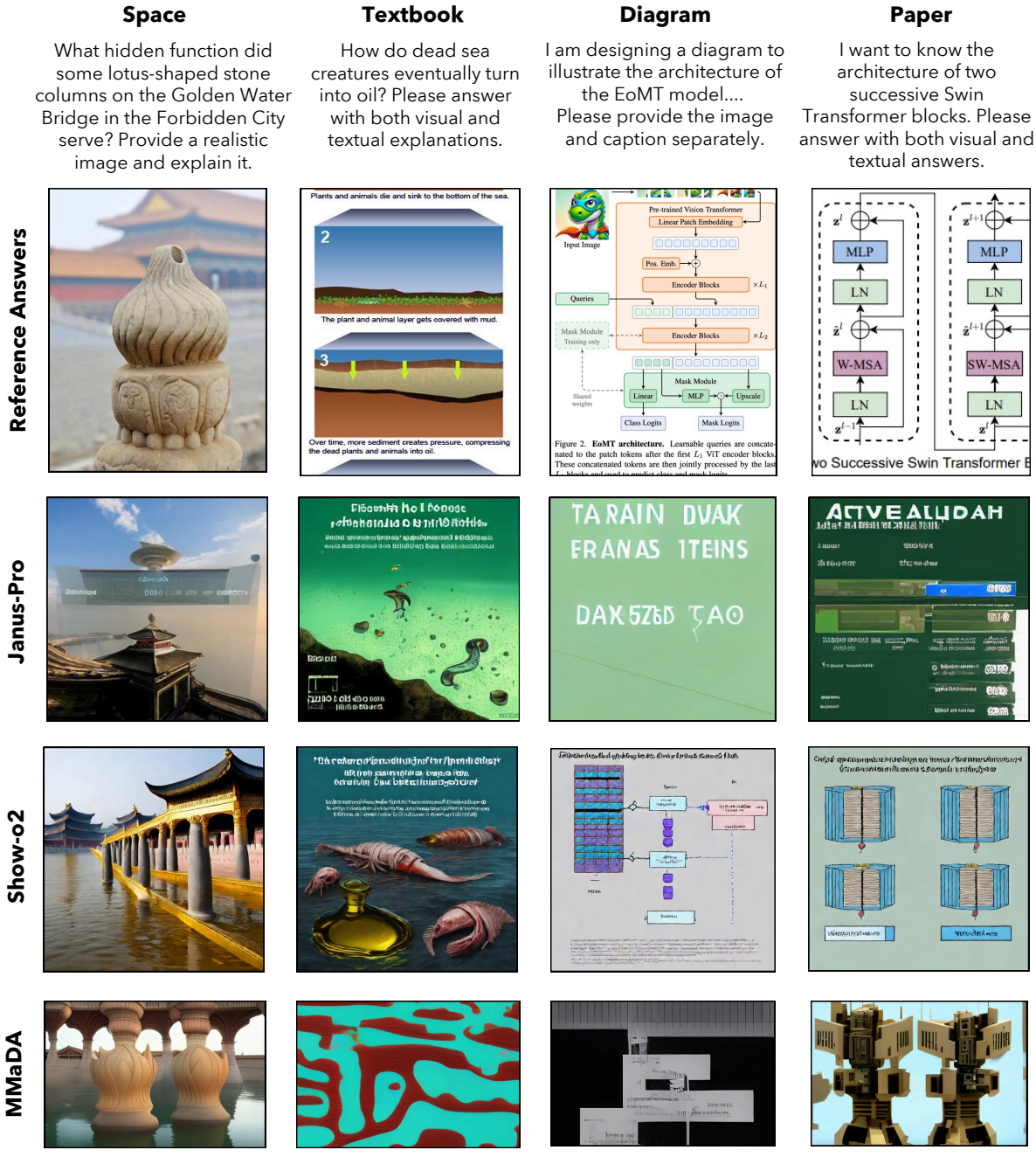

Figure 11 Model-generated images for closed-ended tasks in UEval. We visualize images generated by Janus-Pro, Show-o2, and MMaDA.

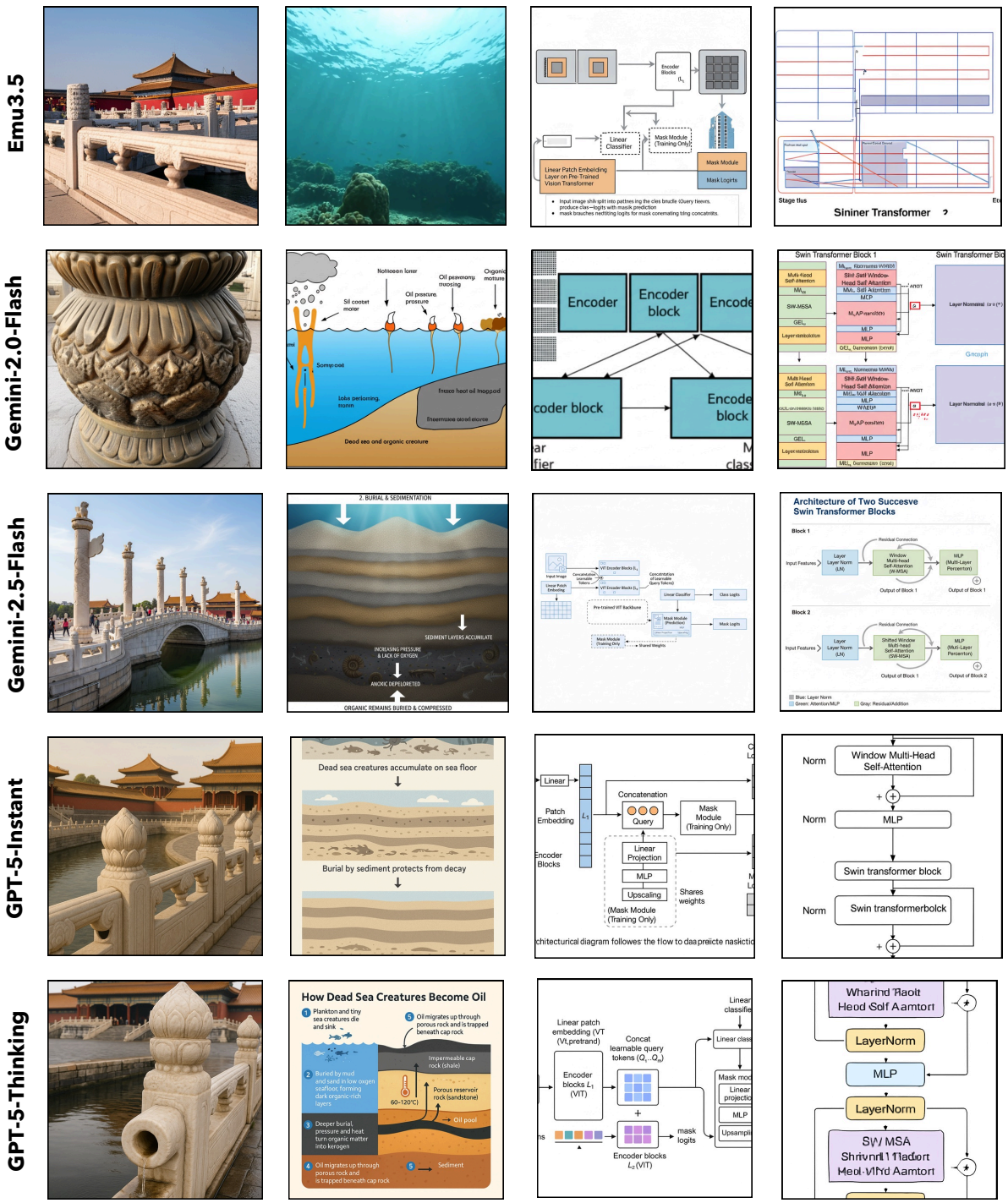

Figure 12 Model-generated images for closed-ended tasks in UEval. We visualize images generated by Emu3.5, Gemini-2.0-Flash, Gemini-2.5-Flash, GPT-5-Instant, and GPT-5-Thinking.

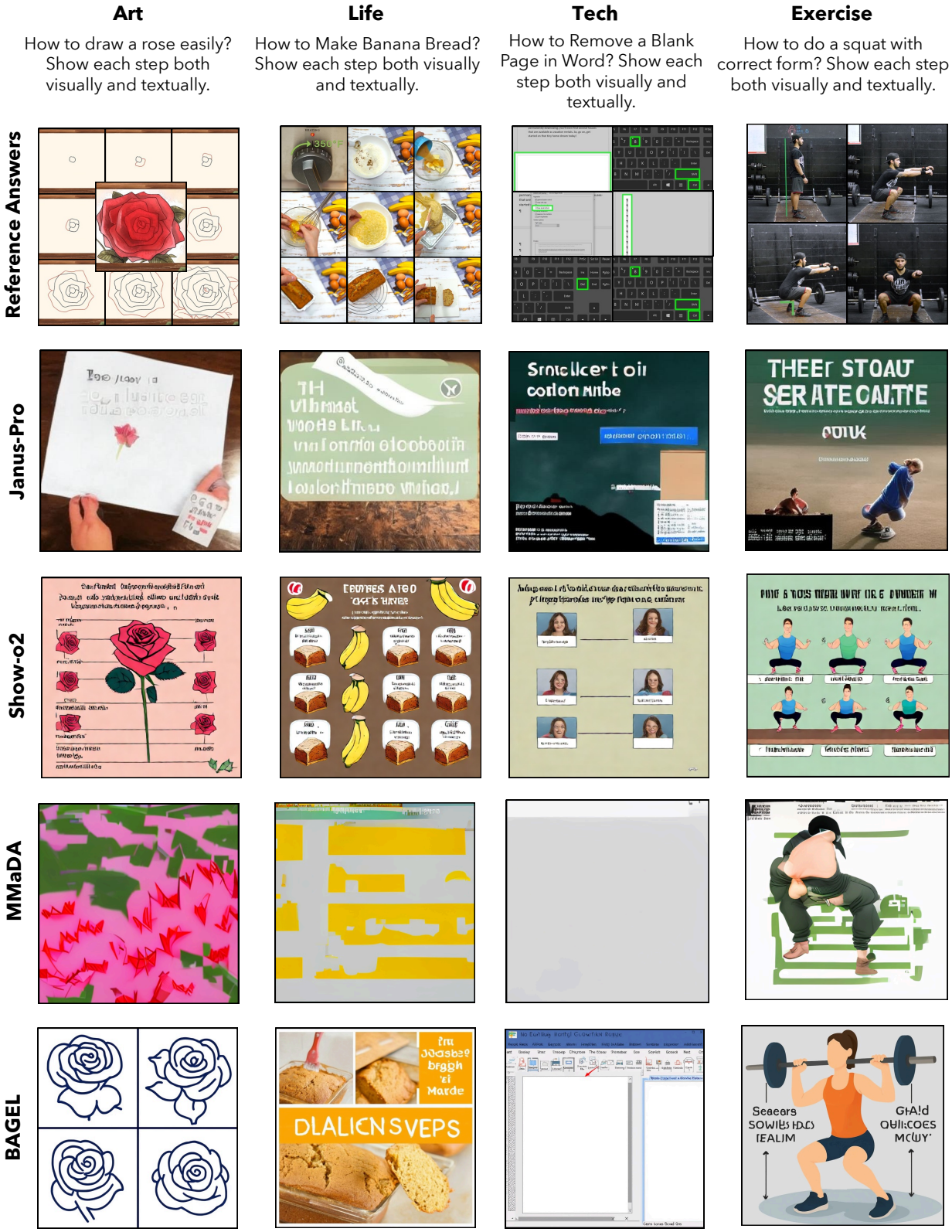

Figure 13 Model-generated images for open-ended tasks in UEval. We prompt Janus-Pro, Show-o2, MMaDA, and BAGEL to produce step-by-step visual guides for each task.

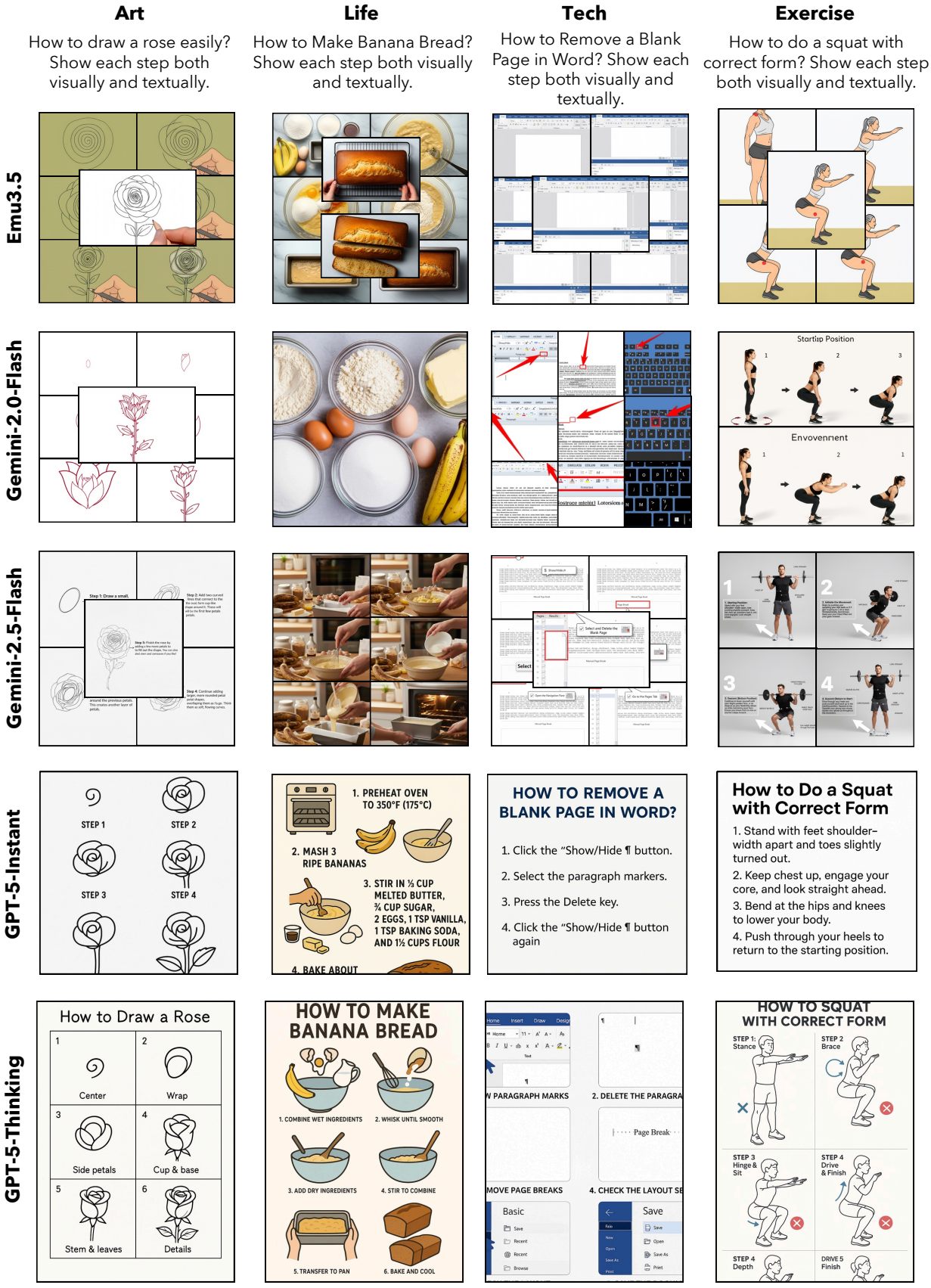

Figure 14 Model-generated images for open-ended tasks in UEval. We prompt Emu3.5, Gemini-2.0-Flash, Gemini-2.5-Flash, GPT-5-Instant, and GPT-5-Thinking to produce step-by-step visual guides for each task.