UEval: A Benchmark for Unified Multimodal Generation

Bo Li Yida Yin Wenhao Chai Xingyu Fu† Zhuang Liu†
† Co-advising

Princeton University

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. Unlike prior methods relying 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 an 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 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 narrows the gap. This suggests that reasoning may be important for tasks requiring complex multimodal understanding and generation. UEval is available at https://huggingface.co/datasets/zlab-princeton/UEval

1 Introduction

Unified multimodal models [14, 53, 67] 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 [17, 36, 38, 63], which requires generating a textual answer from one or more input images, and text-to-image generation [19, 27, 33], 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: unified multimodal generation that *produces both text and images* in response to a single query (Fig. 1). In many real-world tasks, effective responses require images to illustrate concepts while producing text to explain them. Without such evaluation, existing benchmarks fail to capture the interplay between language and vision that characterizes real-world multimodal reasoning.

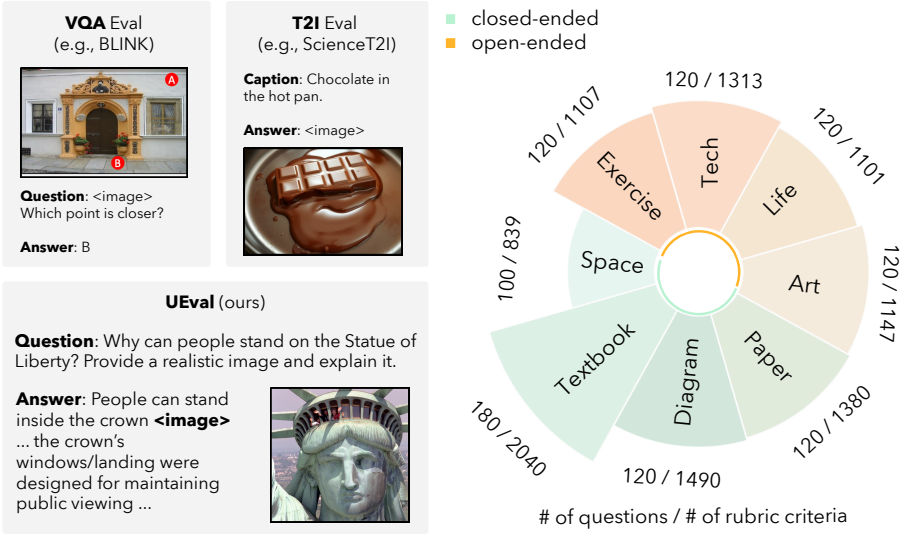

Fig. 1: Left: Previous unified model evaluations focus on either image understanding (*i.e.*, VQA) or image generation from captions (*i.e.*, T2I). In contrast, UEval requires models to reason across modalities and generate responses in both images and text. The VQA example is from BLINK [18], and the T2I example is from ScienceT2I [29]. **Right**: The chart illustrates the number of questions and rubric criteria in UEval.

While recent efforts [2, 35, 41, 60, 66] 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 [8, 23, 56, 61, 62] 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.

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. [3], 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 [22] 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 for all models. Among them, GPT-

5-Thinking [42] achieves the highest score of 66.4 averaged across tasks, whereas the best-performing open-source model (*i.e.*, Emu3.5 [11]) reaches 49.1. We also observe that current models struggle to generate multiple images with consistent labels 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 benefit 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 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 [57], long studied in Large Language Models (LLMs), may also play a role in unified multimodal generation.

2 UEval

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

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

Figures 2 and 3 illustrate the 8 tasks in UEval. 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 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 [22] and then refined by humans (see Section 2.2 for more details). 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.

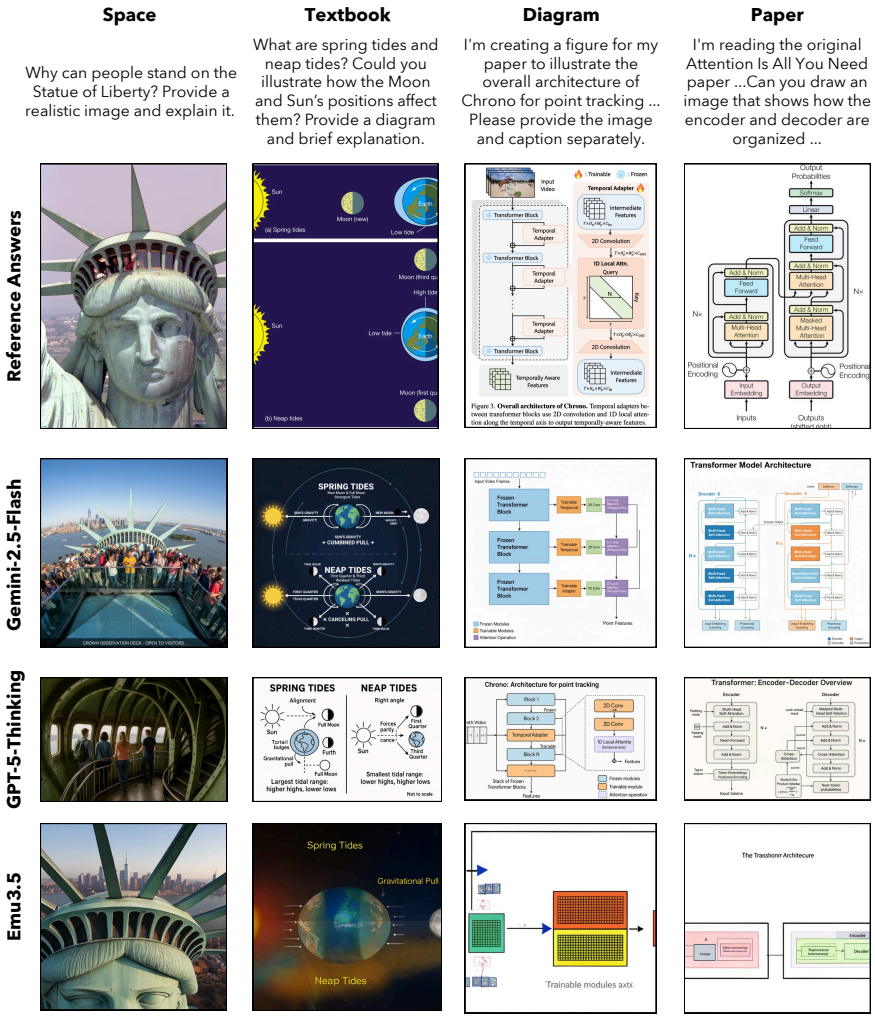

Fig. 2: Model-generated images for closed-ended tasks in UEval. We visualize images generated by GPT-5-Thinking, Gemini-2.5-Flash, and Emu3.5 [11]. These images fail to answer the questions accurately. For example, Gemini-2.5-Flash depicts a nonexistent external platform above the Statue of Liberty instead of the crown interior.

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 set of 20 seed questions about landmarks from online Q&A forums (*e.g.*, Quora-like platforms). Since there are not

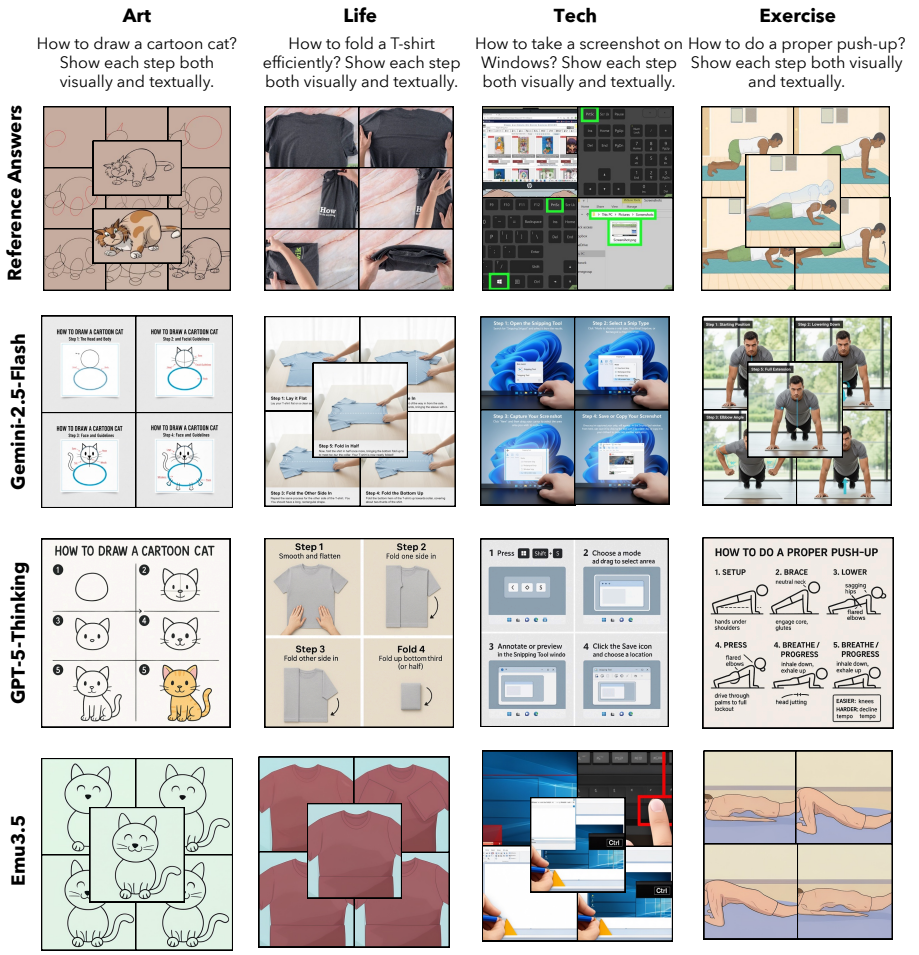

Fig. 3: Model-generated images for open-ended tasks in UEval. We prompt GPT-5-Thinking, Gemini-2.5-Flash, and Emu3.5 to synthesize step-by-step visual guides for each task. The generated images often exhibit temporal inconsistencies. For instance, in the *art* task, when drawing a cartoon cat, GPT-5-Thinking 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.

many such questions online or in datasets, we use GPT-5 to expand this set. Given seed questions, the model is prompted to propose additional questions for landmarks. Human annotators then review generated questions to ensure each refers to real, identifiable features of a landmark.

For every 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 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 [28] 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 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 instructions to check scientific correctness and relevance to the 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.

We source a diverse set of figures from seminal papers (*e.g.*, Transformer [54], ResNet [26]) as well as online teaching platforms (*e.g.*, D2L [64]). 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 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."*

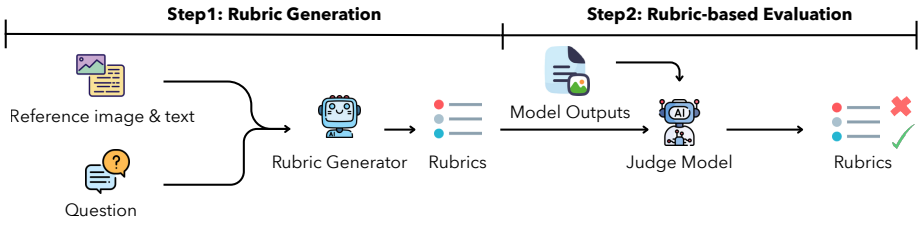

Fig. 4: Rubric drafting and response evaluation procedure in UEval. We propose using *data-dependent* rubrics to evaluate outputs from unified models. For each question, a model drafts an itemized rubric based on the question and the reference image-text pair. A judge model then scores the response against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

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). Because 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 evaluate unified multimodal generation remains an open problem. Zhou et al. [68] 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 model scores. These scores can change if a different set of models is evaluated. Moreover, most current evaluations [60, 68] 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 [3], we adopt a *data-dependent* approach to evaluate unified multimodal generation. Fig. 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. Fig. 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"*.

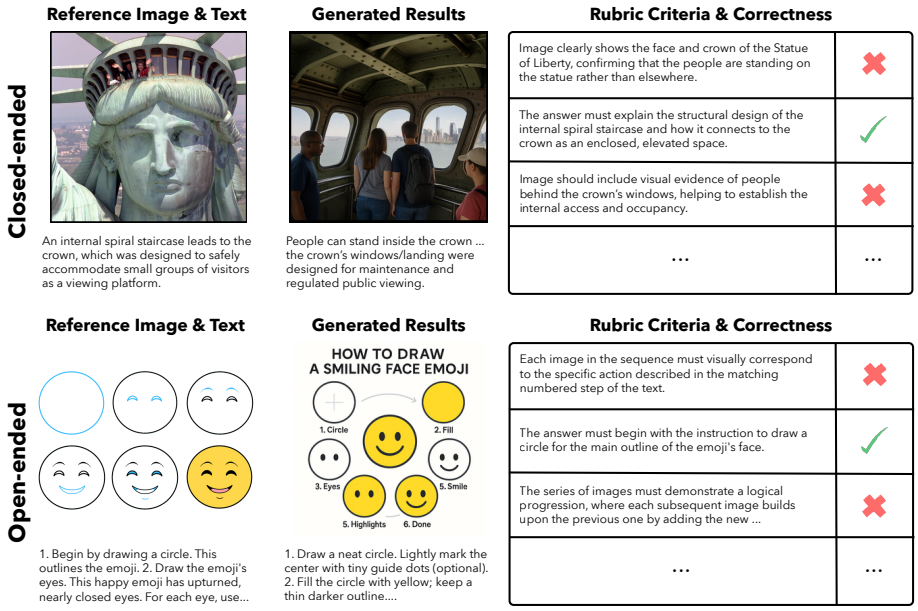

Fig. 5: Rubric examples for closed-ended and open-ended tasks in UEval. Rubrics for closed-ended tasks check whether generated outputs contain specific details important for answering the questions, whereas rubrics for open-ended tasks evaluate higher-level generation qualities (*e.g.*, image-text alignment).

Human review. We conduct two rounds of human review to ensure the quality and reliability of all generated 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 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. For open-source models, we consider Janus-Pro [9], Show-o2 [61], MMaDA [62], BAGEL [14], and Emu3.5 [11]. For proprietary frontier models, we evaluate Gemini-2.0-Flash [21], Gemini-2.5-Flash (*i.e.*, Nano Banana) [23], GPT-5-Instant [42], and GPT-5-Thinking [42]. 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 [22] to grade the generated outputs based on fine-grained rubrics (Section 2.2). Appendix provides the full prompt to evaluate model responses.

3.2 Results

Table 1 reports the performance of various models. 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, Fig. 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.

We also observe that tasks requiring multi-step planning (*e.g.*, art, life) yield substantially lower scores than knowledge-based tasks (*e.g.*, textbook, diagram).

	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).

Fig. 3 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 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. Interestingly, current *reasoning* models (*e.g.*, GPT-5-Thinking) achieve 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.

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 models, their text generation ability lags behind that of text-only models, indicating that high-quality multimodal generation remains challenging.

	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. We also report the text scores of text-only models for comparison.

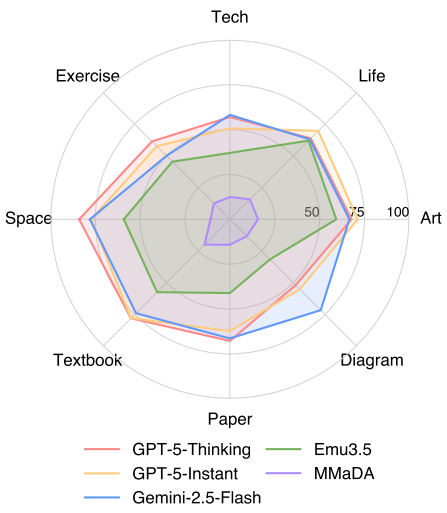

Fig. 6: Model performance on UEval. Each axis corresponds to a task, and each colored polygon is a different model. Proprietary frontier models (*e.g.*, GPT-5-Thinking) consistently outperform open-source models (*e.g.*, Emu3.5).

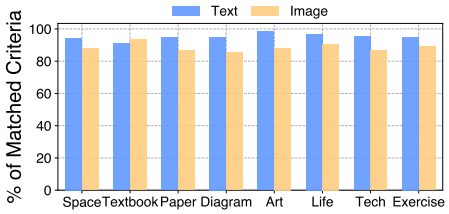

Fig. 7: Percentage of matched rubric criteria between human evaluators and an MLLM judge. We find high agreement between them.

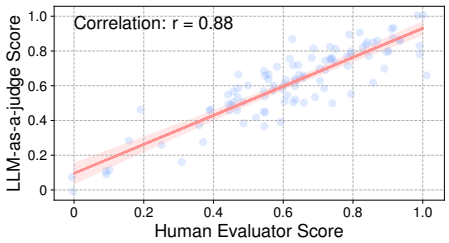

Fig. 8: LLM-as-a-judge *vs.* human evaluator. Each point represents a score for a question in UEval.

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 Fig. 7, we report the percentage of rubric criteria on which human evaluators and the LLM-as-a-judge agree. The results show that, across tasks, approximately 90% of rubric criteria are matched between human evaluators and the LLM-as-a-judge. LLM-as-a-judge grading aligns very closely with human judgment across different tasks. Moreover, in Fig. 8, we plot LLM-as-a-judge scores against human evaluator scores for each question and observe a high Pearson correlation (r = 0.88). These results indicate that our automated evaluation framework is robust and closely aligned with human judgment.

Different judge models. Our default judge model in Table 1 is Gemini-2.5-Pro [22]. 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 [42], Seed1.6-Vision [48], Qwen3-VL-235B-Thinking/Instruct [4], and GLM-4.1V-Thinking [20].

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

	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.

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. Fig. 9 visualizes images generated by some non-reasoning models. Surprisingly, incorporating the reasoning trace enables GPT-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 [57] 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 [1, 12, 30, 34, 69] typically integrate a visual encoder [16, 25, 46] with a pre-trained Large Language Model [37, 43], achieving strong performance on image captioning [10, 44] and visual question answering [24, 40]. To further scale their capabilities, these models require millions of instruction-tuning data [13, 52].

A parallel line of research seeks to unify multimodal understanding and generation within a single model. Some methods adopt diffusion-based approaches [15, 32, 49, 55, 62], whereas others train models with a purely autoregressive objective [45, 51, 56, 58, 59]. There are also hybrid methods that combine both approaches [7, 14, 61, 67]. We refer readers to Zhang et al. [65] for a comprehensive survey of MLLMs and unified models.

Multimodal benchmarks. A range of benchmarks has been proposed to evaluate multimodal inputs. Initial work [24, 38, 39] evaluates image understanding for specific image types, and later efforts benchmark broader image coverage [36, 63]. There are also studies evaluating text-to-image generation quality [19, 27, 33, 47].

Fig. 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 often steer the generated images toward that of the reasoning model and lead to more accurate responses.

Some benchmarks (*e.g.*, VDC [5]) begin to use data-dependent rubrics for better evaluation. More recent works unify understanding and generation benchmarks as interleaved text-and-image generation [2, 6, 35] or unified multimodal generation [31, 50, 70]. 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 [68] or training a scoring model [60] 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 research on developing stronger models and new benchmarks for multimodal generation.

Acknowledgements

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.

Bibliography

[1] Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. In: NeurIPS (2022)

[2] An, J., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, L., Luo, J.: Openleaf: A novel benchmark for open-domain interleaved image-text generation. In: ACM MM (2024)

[3] Arora, R.K., Wei, J., Hicks, R.S., Bowman, P., Quiñonero-Candela, J., Tsimpourlas, F., Sharman, M., Shah, M., Vallone, A., Beutel, A., Heidecke, J., Singhal, K.: Healthbench: Evaluating large language models towards improved human health. arXiv preprint arXiv:2505.08775 (2025)

[4] Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)

[5] Chai, W., Song, E., Du, Y., Meng, C., Madhavan, V., Bar-Tal, O., Hwang, J.N., Xie, S., Manning, C.D.: Auroracap: Efficient, performant video detailed captioning and a new benchmark. In: ICLR (2025)

[6] Chen, D., Chen, R., Pu, S., Liu, Z., Wu, Y., Chen, C., Liu, B., Huang, Y., Wan, Y., Zhou, P., et al.: Interleaved scene graphs for interleaved text-and-image generation assessment. In: ICLR (2025)

[7] Chen, J., Xu, Z., Pan, X., Hu, Y., Qin, C., Goldstein, T., Huang, L., Zhou, T., Xie, S., Savarese, S., Xue, L., Xiong, C., Xu, R.: Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. arXiv preprint arXiv:2505.09568 (2025)

[8] Chen, L., Bai, S., Chai, W., Xie, W., Zhao, H., Vinci, L., Lin, J., Chang, B.: Multimodal representation alignment for image generation: Text-image interleaved control is easier than you think. arXiv preprint arXiv:2502.20172 (2025)

[9] Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025)

[10] Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)

[11] Cui, Y., Chen, H., Deng, H., Huang, X., Li, X., Liu, J., Liu, Y., Luo, Z., Wang, J., Wang, W., Wang, Y., Wang, C., Zhang, F., Zhao, Y., Pan, T., Li, X., Hao, Z., Ma, W., Chen, Z., Ao, Y., Huang, T., Wang, Z., Wang,

X.: Emu3.5: Native multimodal models are world learners. arXiv preprint arXiv:2510.26583 (2025)
[12] Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. In: NeurIPS (2023)
[13] Deitke, M., Clark, C., Lee, S., Tripathi, R., Yang, Y., Park, J.S., Salehi, M., Muennighoff, N., Lo, K., et al.: Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In: CVPR (2025)
[14] Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
[15] Dong, R., Han, C., Peng, Y., Qi, Z., Ge, Z., Yang, J., Zhao, L., Sun, J., Zhou, H., Wei, H., Kong, X., Zhang, X., Ma, K., Yi, L.: Dreamllm: Synergistic multimodal comprehension and creation. In: ICLR (2024)
[16] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
[17] Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., Ji, R.: Mme: A comprehensive evaluation benchmark for multimodal large language models. In: NeurIPS (2025)
[18] Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.C., Krishna, R.: Blink: Multimodal large language models can see but not perceive. In: ECCV (2024)
[19] Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment. In: NeurIPS (2023)
[20] GLM: Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. arXiv preprint arXiv:2507.01006 (2025)
[21] 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)
[22] Google: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv: 2507.06261 (2025)
[23] 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/ (2025)
[24] Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: CVPR (2017)
[25] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR (2022)
[26] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)

[27] Huang, K., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In: NeurIPS (2023)
[28] Kembhavi, A., Seo, M., Schwenk, D., Choi, J., Farhadi, A., Hajishirzi, H.: Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In: CVPR (2017)
[29] Li, J., Chai, W., Fu, X., Xu, H., Xie, S.: Science-t2i: Addressing scientific illusions in image synthesis. In: CVPR (2025)
[30] Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML (2023)
[31] Li, Y., Wang, H., Zhang, Q., Xiao, B., Hu, C., Wang, H., Li, X.: Unieval: Unified holistic evaluation for unified multimodal understanding and generation. arXiv preprint arXiv:2505.10483 (2025)
[32] Li, Z., Li, H., Shi, Y., Farimani, A.B., Kluger, Y., Yang, L., Wang, P.: Dual diffusion for unified image generation and understanding. In: CVPR (2025)
[33] Lin, Z., Pathak, D., Li, B., Li, J., Xia, X., Neubig, G., Zhang, P., Ramanan, D.: Evaluating text-to-visual generation with image-to-text generation. In: ECCV (2024)
[34] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
[35] Liu, M., Xu, Z., Lin, Z., Ashby, T., Rimchala, J., Zhang, J., Huang, L.: Holistic evaluation for interleaved text-and-image generation. In: EMNLP (2024)
[36] Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: ECCV (2024)
[37] Llama-2-Team: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
[38] Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: CVPR (2019)
[39] Masry, A., Do, X.L., Tan, J.Q., Joty, S., Hoque, E.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In: ACL (2022)
[40] Mathew, M., Karatzas, D., Jawahar, C.V.: Docvqa: A dataset for vqa on document images. In: WACV (2021)
[41] Niu, Y., Ning, M., Zheng, M., Jin, W., Lin, B., Jin, P., Liao, J., Feng, C., Ning, K., Zhu, B., et al.: Wise: A world knowledge-informed semantic evaluation for text-to-image generation. arXiv preprint arXiv:2503.07265 (2025)
[42] OpenAI: Introducing gpt-5. https://openai.com/index/introducing-gpt-5/ (2025)
[43] Peng, B., Li, C., He, P., Galley, M., Gao, J.: Instruction tuning with gpt-4. arXiv preprint arXiv:2304.03277 (2023)
[44] Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: ICCV (2015)

[45] Qu, L., Zhang, H., Liu, Y., Wang, X., Jiang, Y., Gao, Y., Ye, H., Du, D.K., Yuan, Z., Wu, X.: Tokenflow: Unified image tokenizer for multimodal understanding and generation. In: CVPR (2025)
[46] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021)
[47] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. In: NeurIPS (2022)
[48] Seed: Seed1.6. https://seed.bytedance.com/en/seed1_6 (2025)
[49] Shi, Q., Bai, J., Zhao, Z., Chai, W., Yu, K., Wu, J., Song, S., Tong, Y., Li, X., Li, X., et al.: Muddit: Liberating generation beyond text-to-image with a unified discrete diffusion model. arXiv preprint arXiv:2505.23606 (2025)
[50] Shi, Y., Dong, Y., Ding, Y., Wang, Y., Zhu, X., Zhou, S., Liu, W., Tian, H., Wang, R., Wang, H., Liu, Z., Zeng, B., Chen, R., Wang, Q., Zhang, Z., Chen, X., Tong, C., Li, B., Fu, C., Liu, Q., Wang, H., Yang, W., Zhang, Y., Wan, P., Zhang, Y.F., Liu, Z.: Realunify: Do unified models truly benefit from unification? a comprehensive benchmark. arXiv preprint arXiv:2509.24897 (2025)
[51] Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024)
[52] Tong, P., Brown, E., Wu, P., Woo, S., IYER, A.J.V., Akula, S.C., Yang, S., Yang, J., Middepogu, M., Wang, Z., et al.: Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In: NeurIPS (2024)
[53] Tong, S., Fan, D., Zhu, J., Xiong, Y., Chen, X., Sinha, K., Rabbat, M., LeCun, Y., Xie, S., Liu, Z.: Metamorph: Multimodal understanding and generation via instruction tuning. In: ICCV (2024)
[54] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
[55] Wang, J., Lai, Y., Li, A., Zhang, S., Sun, J., Kang, N., Wu, C., Li, Z., Luo, P.: Fudoki: Discrete flow-based unified understanding and generation via kinetic-optimal velocities. arXiv preprint arXiv:2505.20147 (2025)
[56] Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024)
[57] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models. In: NeurIPS (2022)
[58] Wu, C., Chen, X., Wu, Z., Ma, Y., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C., et al.: Janus: Decoupling visual encoding for unified multimodal understanding and generation. In: CVPR (2025)
[59] Wu, Y., Zhang, Z., Chen, J., Tang, H., Li, D., Fang, Y., Zhu, L., Xie, E., Yin, H., Yi, L., Han, S., Lu, Y.: Vila-u: a unified foundation model integrating visual understanding and generation. In: ICLR (2025)

[60] Xia, P., Han, S., Qiu, S., Zhou, Y., Wang, Z., Zheng, W., Chen, Z., Cui, C., Ding, M., Li, L., et al.: Mmie: Massive multimodal interleaved comprehension benchmark for large vision-language models. In: ICLR (2025)
[61] Xie, J., Yang, Z., Shou, M.Z.: Show-o2: Improved native unified multimodal models. In: NeurIPS (2025)
[62] Yang, L., Tian, Y., Li, B., Zhang, X., Shen, K., Tong, Y., Wang, M.: Mmada: Multimodal large diffusion language models. In: NeurIPS (2025)
[63] Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., Sun, H., Su, Y., Chen, W.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: CVPR (2024)
[64] Zhang, A., Lipton, Z.C., Li, M., Smola, A.J.: Dive into deep learning. https://d2l.ai (2023)
[65] Zhang, X., Guo, J., Zhao, S., Fu, M., Duan, L., Hu, J., Chng, Y.X., Wang, G.H., Chen, Q.G., Xu, Z., Luo, W., Zhang, K.: Unified multimodal understanding and generation models: Advances, challenges, and opportunities. arXiv preprint arXiv:2505.02567 (2025)
[66] Zhao, X., Zhang, P., Tang, K., Zhu, X., Li, H., Chai, W., Zhang, Z., Xia, R., Zhai, G., Yan, J., et al.: Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing. In: NeurIPS (2025)
[67] Zhou, C., Yu, L., Babu, A., Tirumala, K., Yasunaga, M., Shamis, L., Kahn, J., Ma, X., Zettlemoyer, L., Levy, O.: Transfusion: Predict the next token and diffuse images with one multi-modal model. In: ICLR (2025)
[68] Zhou, P., Peng, X., Song, J., Li, C., Xu, Z., Yang, Y., Guo, Z., Zhang, H., Lin, Y., He, Y., et al.: Opening: A comprehensive benchmark for judging open-ended interleaved image-text generation. In: CVPR (2025)
[69] Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. In: ICLR (2024)
[70] Zou, K., Huang, Z., Dong, Y., Tian, S., Zheng, D., Liu, H., He, J., Liu, B., Qiao, Y., Liu, Z.: Uni-mmmu: A massive multi-discipline multimodal unified benchmark. arXiv preprint arXiv:2510.13759 (2025)