# GRADE: Benchmarking Discipline-Informed Reasoning in Image Editing

Mingxin Liu[1*], Ziqian Fan[2*], Zhaokai Wang[1*‡], Leyao Gu[1*], Zirun Zhu[1*], Yiguo He[1], Yuchen Yang[3], Changyao Tian[4], Xiangyu Zhao[1], Ning Liao[1], Shaofeng Zhang[5], Qibing Ren[1], Zhihang Zhong[1], Xuanhe Zhou[1], Junchi Yan[1], and Xue Yang[1†]

[1] Shanghai Jiao Tong University
[2] South China University of Technology
[3] Fudan University
[4] The Chinese University of Hong Kong
[5] University of Science and Technology of China

*Equal contribution ‡Project lead †Corresponding author

Project Page: https://grade-bench.github.io/
Evaluation Code: https://github.com/VisionXLab/GRADE
Benchmark: https://huggingface.co/datasets/VisionXLab/GRADE

**Abstract.** Unified multimodal models target joint understanding, reasoning, and generation, but current image editing benchmarks are largely confined to natural images and shallow commonsense reasoning, offering limited assessment of this capability under structured, domain-specific constraints. In this work, we introduce GRADE, the first benchmark to assess discipline-informed knowledge and reasoning in image editing. GRADE comprises 520 carefully curated samples across 10 academic domains, spanning from natural science to social science. To support rigorous evaluation, we propose a multi-dimensional evaluation protocol that jointly assesses Discipline Reasoning, Visual Consistency, and Logical Readability. Extensive experiments on 20 state-of-the-art open-source and closed-source models reveal substantial limitations in current models under implicit, knowledge-intensive editing settings, leading to large performance gaps. Beyond quantitative scores, we conduct rigorous analyses and ablations to expose model shortcomings and identify the constraints within disciplinary editing. Together, GRADE pinpoints key directions for the future development of unified multimodal models, advancing the research on discipline-informed image editing and reasoning. Our benchmark and evaluation code are publicly released.

**Keywords:** Benchmark, Image Editing, Unified Multimodal Models

## 1 Introduction

Recent years have witnessed rapid advances in unified multimodal models (UMMs), *i.e.* models with unified abilities of multimodal understanding and image generation [6,10,23,32]. Modern UMMs are expected to integrate knowledge, structured

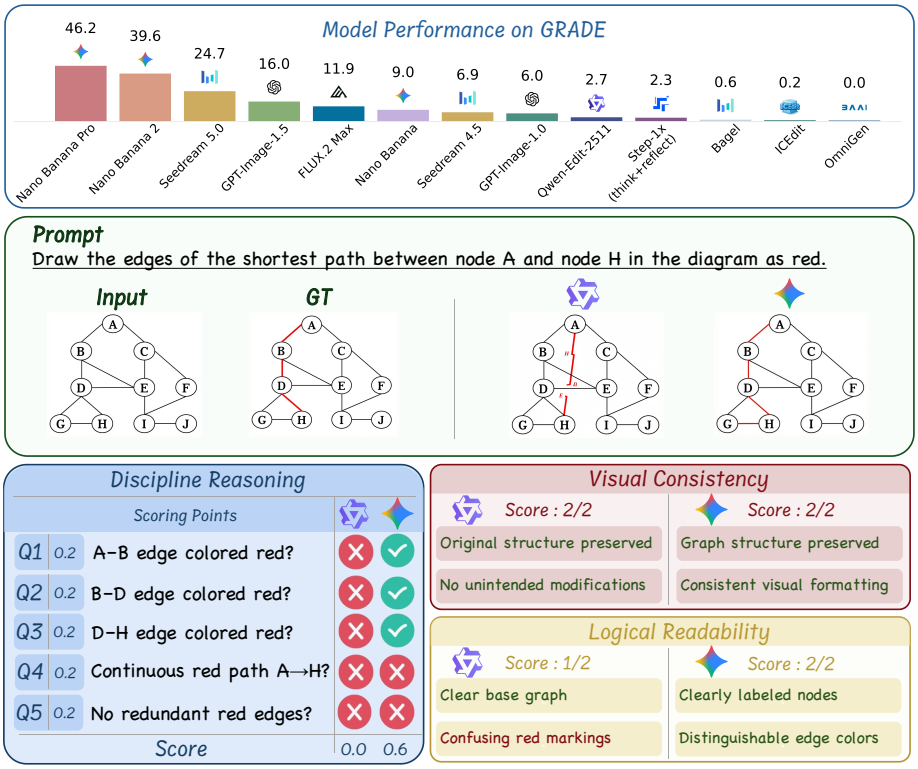


**Fig. 1: Examples of images generated by state-of-the-art editing models on GRADE and their evaluation results.** Challenging discipline-informed image editing exposes limitations of current models in complex knowledge reasoning. Notable performance gaps exist across models on GRADE.

reasoning, and controllable generation within a single system. In this context, several existing benchmarks [31,47] attempt to evaluate reasoning in image editing. They are predominantly grounded in natural image domains, where reasoning difficulty mainly arises from the implicitness or linguistic complexity of prompts, and the underlying knowledge involved is typically shallow, relying largely on everyday commonsense. As a result, these settings provide only a limited test of whether unified multimodal models can truly coordinate knowledge, reasoning, and editing in a principled manner.

Inspired by prior work that introduces discipline-specific knowledge as a challenging axis in multimodal understanding [24, 41] and text-to-image generation [20, 34] tasks, we argue that discipline-informed image editing offers a more stringent and representative evaluation setting, as it requires models to operate over structured discipline-informed knowledge, which is inherently more demanding than everyday commonsense. In contrast to prior evaluation settings

that assess discipline-specific knowledge grounding either through pure image understanding or unconditioned image generation, image editing requires models to reason under domain constraints while preserving existing visual structures and making precise modifications, suggesting greater challenges for UMMs.

Motivated by this gap, we introduce GRADE, which stands for **G**rounded **R**easoning **A**ssessment for **D**iscipline-informed **E**diting. GRADE is the first image editing benchmark explicitly designed to evaluate discipline-informed reasoning. While prior benchmarks mainly emphasize the implicitness or linguistic complexity of prompts, GRADE enriches knowledge and reasoning difficulty through the challenging discipline-specific image domain. GRADE comprises 520 carefully curated samples spanning 10 academic domains, from natural sciences to humanities and applied fields. They reflect practical image editing scenarios such as assisting researchers or teaching assistants in correcting geometric diagrams, modifying chemical structures, or refining data visualizations, thereby providing a comprehensive and challenging testbed for evaluating discipline-informed reasoning in image editing. For the evaluation metrics, as shown in Fig. 1, we introduce Discipline Reasoning for explicitly measuring discipline-informed knowledge reasoning with fine-grained scoring points, Visual Consistency for task-dependent consistency constraints, and Logical Readability for evaluating logically clear and well-structured academic representations. These dimensions provide a comprehensive evaluation framework that goes beyond aesthetic quality and realism in general image editing to assess the rigor and interpretability of discipline-informed image editing.

We evaluate 20 state-of-the-art (SoTA) image editing models on GRADE. We observe that models with comparable performance on other benchmarks, such as Nano Banana Pro [10] and GPT-Image-1.5 [23], exhibit markedly different behaviors under discipline-informed editing (46.2% vs. 16.0%), highlighting GRADE's stronger discriminative power in probing implicit academic knowledge and structured reasoning. We also observe a huge gap between open-source and closed-source models, where the SoTA open-source model Qwen-Edit-2511 [35] (2.7%) significantly lags behind closed-source SoTA.

Overall, implicit discipline-informed reasoning remains a major bottleneck in current models. To further examine this limitation, we analyze representative failure cases and compare implicit and explicit instruction formulations, providing additional diagnostic insights into current model behavior. In sum, GRADE provides clear and actionable directions for the future advancement of UMMs towards discipline-informed image editing and reasoning tasks.

Overall, our contributions are as follows:

1. We present the first discipline-informed image editing benchmark spanning multiple academic domains, designed to rigorously evaluate knowledge grounding and reasoning in image editing models.

2. We propose a multi-dimensional automated evaluation pipeline that is scalable and exhibits strong alignment with human judgments.

3. We conduct comprehensive analysis across a wide range of SoTA models, revealing key limitations of current approaches and offering actionable guidance for future research.

## 2 Related Work

### 2.1 Image Generation Models

Recent advances in image generation are primarily driven by diffusion-based models [7,13], which serve as the foundation for modern text-conditioned synthesis. Early text-to-image systems, such as VQGAN-based pipelines and cascaded diffusion models [3,5], established the feasibility of aligning visual synthesis with natural language descriptions, while subsequent works improved photorealism and semantic faithfulness through large-scale training and refined conditioning mechanisms [25, 44].

More recently, a line of research has shifted toward unified generative frameworks that aim to integrate language understanding, visual perception, and image generation within a single model [6, 17, 32, 42]. Open-source efforts such as OmniGen [38] emphasize accessibility and extensibility, enabling community-driven exploration of unified generation capabilities. Meanwhile, large-scale proprietary systems, including Gemini [33], Seedream [29], and recent OpenAI models [30], demonstrate strong visual quality and robustness through extensive data and infrastructure support.

In addition, image editing models [18] such as Step-1X [19], Qwen-Image [35], and others [45] further extend text-to-image frameworks by incorporating an input image and a textual instruction to perform targeted modifications. While recent models often produce visually plausible edits, their ability to handle reasoning-intensive scenarios, particularly those requiring structured, domain-specific knowledge, remains largely unexplored.

### 2.2 Image Editing Benchmarks

To enable systematic evaluation of generative models, a range of benchmarks [15, 39] have been proposed to assess visual quality [12, 26, 40] and semantic alignment [8, 14]. However, most existing image editing benchmarks rely on highly explicit instructions that directly specify the required operations. ImgEdit [39] primarily targets traditional editing tasks where the required operations are explicitly specified and reasoning is not a central concern. As UMMs rapidly advance, the community has placed increasing emphasis on evaluating their reasoning capabilities. More recent editing benchmarks begin to incorporate implicit reasoning, but the involved knowledge is typically restricted to general-purpose commonsense rather than disciplinary expertise. RISEBench [47] categorizes reasoning into temporal, causal, spatial, and logical types, while KRISBench [36] organizes reasoning knowledge based on cognitively motivated taxonomies. While these benchmarks effectively assess general reasoning, they do not focus on the assessment of disciplinary knowledge in editing.

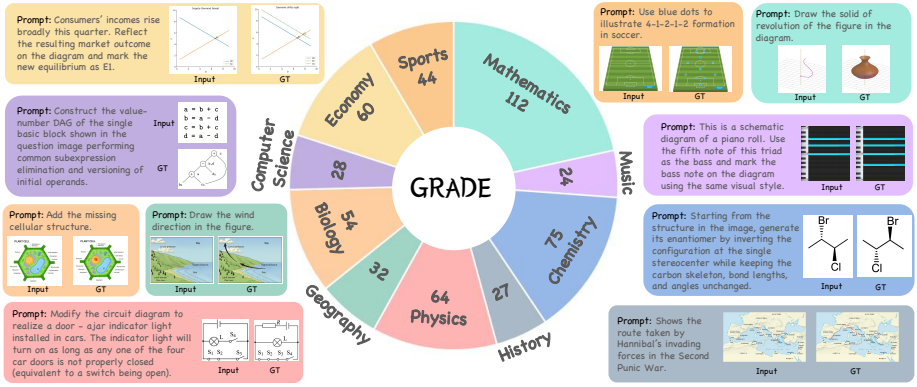


**Fig. 2: Overview of GRADE.** GRADE contains 520 discipline-informed image editing samples across ten academic disciplines.

### 2.3 Discipline-Specific Benchmarks

Disciplinary knowledge is widely regarded as one of the most challenging forms of reasoning and has been extensively explored in previous benchmarks. In multimodal understanding, MMMU [41] evaluates multimodal reasoning across six broad discipline categories spanning over thirty disciplines. Similarly, HLE [24] includes image-related tasks that require PhD-level, multi-disciplinary understanding, emphasizing high-difficulty academic reasoning in visual comprehension. In the text-to-image setting, MMMG [20] and Sridbench [4] focus on disciplinary concept illustration, and GenExam [34] further explores generation tasks grounded in professional disciplinary knowledge. Despite these advances, disciplinary reasoning remains largely unexplored in the image editing setting, which requires tighter integration of knowledge understanding, reasoning, and editing capabilities.

## 3 GRADE Benchmark

### 3.1 Overview

**Data Source.** Each sample in our dataset is organized as an image editing triplet, consisting of an input image, a textual editing instruction, and a corresponding GT image. Given the specialized and technically demanding nature of disciplinary knowledge, we adopt a carefully designed data construction and filtering pipeline to ensure both rigor and reliability. For the majority of the dataset, six annotators with academic backgrounds in relevant disciplines source concept-grounded images from open textbooks, websites and other reference materials, then manually edit them to create input-GT pairs and design corresponding instructions, followed by cross-validation by two additional experts. For the

remainder, we first apply an automated pipeline to coarsely filter suitable samples from MMMU [41], followed by manual curation where two experts select final samples and design editing instructions. The resulting data then undergoes cross-validation by two additional experts. Additional details on data collection are provided in the supplementary material.

**Taxonomy.** Fig. 2 illustrates the data distribution of our dataset along with representative examples. Our dataset covers 10 academic disciplines, including *mathematics*, *physics*, *chemistry*, *biology*, *history*, *geography*, *sports*, *music*, *computer science*, and *economics*. To further capture fine-grained knowledge structures, we introduce a hierarchical categorization scheme by defining second-level sub-disciplines within each primary domain. For example, within the mathematics domain, we distinguish sub-disciplines such as *plane geometry*, *solid geometry*, *functions*, *graph* and *statistics*, each targeting distinct forms of domain-specific knowledge and reasoning patterns. Similar hierarchical decompositions are applied across other disciplines to better reflect the diversity of academic concepts encountered in real-world editing scenarios. The detailed sub-discipline taxonomy for all ten disciplines is provided in the supplementary material.

### 3.2 Evaluation Metrics

As the reasoning difficulty shifts toward complex discipline-informed knowledge, the design and rigor of evaluation become increasingly critical. To this end, we evaluate image editing performance along three dimensions: *Discipline Reasoning*, *Visual Consistency*, and *Logical Readability*. Together, these complementary dimensions provide a comprehensive assessment of editing quality, covering both reasoning correctness and low-level visual quality. The overall evaluation pipeline is illustrated in Fig. 3.

**Discipline Reasoning.** We evaluate reasoning performance by assessing whether edited results correctly reflect the underlying disciplinary knowledge, which is central to the validity of discipline-informed image editing. Relying solely on human evaluation is costly and difficult to scale, while MLLM-as-a-judge [46] approaches based on a single instruction often lead to unstable judgments. In practice, such prompts may inadequately capture all required criteria, resulting in high sensitivity to phrasing and limited reproducibility across evaluations.

Therefore, inspired by previous approaches [34, 43], we adopt a structured, question-guided evaluation strategy. For each sample, we use GPT-5 [30] to generate weighted binary questions aligned with required disciplinary knowledge, with all weights summing to one. Two human experts with relevant disciplinary knowledge independently examine all scoring points, and the results are cross-validated by a third expert. During evaluation, we use Gemini-3-Flash [9] to assess the edited result by explicitly referencing these scoring points together with the GT image. The final reasoning score for each sample is computed as a weighted aggregation of the judge's responses, yielding a normalized score between 0 and 1. By decomposing discipline-informed reasoning into explicit,

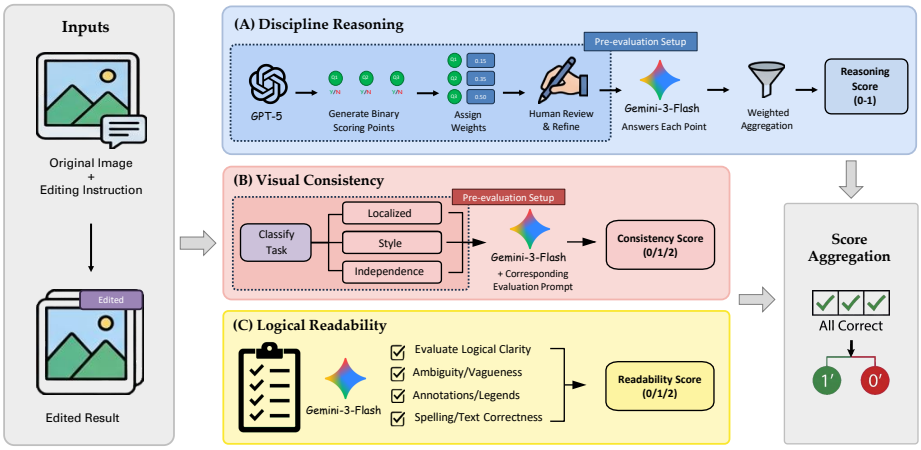


**Fig. 3: Evaluation pipeline.** We evaluate edited results on (A) *Discipline Reasoning* via weighted, question-guided MLLM judging, (B) *Visual Consistency* with task-specific prompts (localized/style/independence), and (C) *Logical Readability* for clarity and text/annotation correctness.

verifiable criteria, the question-guided protocol enables targeted evaluation of whether models correctly understand and apply disciplinary concepts.

**Visual Consistency.** Visual consistency is a fundamental aspect of image editing quality, as it reflects whether an edit integrates coherently with the intended visual structure [36, 47]. However, consistency requirements are inherently task-specific and vary across different editing scenarios. To this end, we categorize consistency evaluation into three types:

1. *Localized Consistency.* This type applies to localized editing tasks, where only specific regions or elements are expected to change. For example, when completing missing entries in a timeline or adjusting curves in an economics diagram, all unrelated visual elements (*i.e.* elements not mentioned in the editing instruction) should remain unchanged. Consistency is satisfied only if modifications are confined to the target components.

2. *Style Consistency.* This type applies to global edits where the overall structure is modified, which can be viewed as style-consistent image-to-image generation. These tasks do not require local consistency of certain elements, but the visual representation style should be preserved. For instance, when editing a chemical reaction diagram, the resulting molecule is expected to maintain the original bond-line representation rather than switching to alternative styles such as ball-and-stick models.

3. *Consistency Independence.* This category applies to scenarios where consistency with the original image is not required, resembling general image-to-image generation. For example, given a rendered image of a mechanical part, the task may require generating its engineering orthographic views. In such cases,

the edited output is expected to follow domain-specific representation standards rather than preserve visual consistency with the input image.

The taxonomy was finalized through consensus among multiple annotators to minimize subjective bias. Based on these three categories, we design corresponding evaluation prompts to evaluate visual consistency. Each sample is assigned a consistency score of $0/1/2$.

**Logical Readability**. Readability generally refers to whether an edited image can be clearly interpreted by a human viewer. Unlike natural image editing, where perceptual realism and visual sharpness are often sufficient, discipline-informed editing places stronger emphasis on whether knowledge is expressed in a logically coherent and structurally sound manner. An image that appears clear at a perceptual level may still fail to convey correct or interpretable academic meaning if its logical structure is flawed or its representations are ambiguous. For instance, curves in a diagram should be visually distinguishable, accompanied by clear annotations or legends, use correct and consistent textual labels, and follow a coherent representational convention to support reliable interpretation in academic contexts. Accordingly, we introduce a criterion for discipline-informed editing to assess whether the edited image presents discipline-specific content in a clear, logically consistent, and interpretable form. Each edited result is assigned a readability score of $0/1/2$.

**Score Aggregation**. Inspired by prior benchmark designs, we adopt an overall accuracy that requires joint satisfaction of all evaluation criteria [47]. A sample is considered correct only if the model achieves the maximum score in all three dimensions; otherwise, it is counted as a failure. Additionally, we report a relaxed score that summarizes performance via criterion-wise aggregation [21]. The relaxed score is computed as a weighted average of the three evaluation dimensions. We first normalize each dimension to range [0, 100], and assign weights of 0.6, 0.3, and 0.1 to Discipline Reasoning, Visual Consistency, and Logical Readability, respectively.

## 4 Experiment

We conduct extensive experiments on our benchmark using a diverse set of 20 state-of-the-art models, including 10 closed-source models and 10 open-source models. The evaluated models cover both unified multimodal models, such as GPT-Image-1.5 [23] and Nano Banana [9], as well as specialized image editing models, including Qwen-Edit [35] and FLUX [16]. Our automated evaluation uses Gemini-3-Flash [9] as the judge. We systematically analyze model performance across evaluation dimensions, with human alignment and ablation experiments to further interpret model behavior and evaluation reliability.

### 4.1 Main Result

The performance of all evaluated models across individual dimensions and overall scores is summarized in Table 1. Among all the evaluated models, Nano Banana

**Table 1:** Performance comparison of different models across individual evaluation dimensions (normalized to 0-100) and the overall accuracy.

| Model | Reasoning | Consistency | Readability | Accuracy |
|---|---|---|---|---|
| **▼ Closed Source Models** | | | | |
| Nano Banana Pro [10] | **77.5** | **89.5** | 95.8 | **46.2** |
| Nano Banana 2 [11] | 72.6 | 86.4 | **95.9** | 39.6 |
| Seedream 5.0 [28] | 64.1 | 87.5 | 90.6 | 24.7 |
| GPT-Image-1.5 [23] | 54.5 | 82.3 | 90.7 | 16.0 |
| FLUX.2 Max [16] | 47.8 | 67.2 | 68.6 | 11.9 |
| Nano Banana [33] | 42.2 | 75.1 | 82.0 | 9.0 |
| Seedream 4.5 [27] | 41.3 | 55.6 | 82.1 | 6.9 |
| GPT-Image-1.0 [22] | 44.0 | 65.2 | 82.3 | 6.0 |
| FLUX.2 Pro [16] | 38.9 | 55.5 | 70.3 | 4.4 |
| Seedream 4.0 [29] | 32.4 | 53.2 | 77.0 | 3.1 |
| **▼ Open Source Models** | | | | |
| Qwen-Edit-2511 [35] | 18.6 | 45.2 | 52.1 | 2.7 |
| Step-1X (think+reflect) [19] | 19.2 | 57.2 | 66.9 | 2.3 |
| FLUX.2 dev [16] | 23.0 | 56.4 | 69.0 | 2.1 |
| Step-1X (think) [19] | 17.6 | 56.3 | 68.2 | 1.4 |
| DreamOmni [37] | 17.4 | 83.2 | 89.1 | 1.0 |
| Step-1X [19] | 17.3 | 52.8 | 63.7 | 1.0 |
| Bagel [6] | 15.2 | 58.6 | 69.8 | 0.6 |
| Bagel (think) [6] | 15.6 | 54.8 | 67.8 | 0.2 |
| ICEdit [45] | 9.8 | 33.2 | 56.5 | 0.2 |
| OmniGen [38] | 9.7 | 33.6 | 51.6 | 0.0 |

Pro [10] achieves the strongest overall performance, exhibiting clear advantages across all three dimensions with an accuracy of 46.2%. Notably, Nano Banana Pro and Nano Banana 2 significantly surpass other systems such as Seedream 5.0 (24.7%). Despite this margin, current models can only attain an accuracy below 50%, indicating that even the best-performing model fails to satisfy all discipline-informed editing requirements in more than half of the cases. Interestingly, models such as Nano Banana Pro, GPT-Image-1.5, and Seedream 5.0, which achieve comparable performance on existing benchmarks [1, 47], exhibit substantial performance differences on GRADE (46.2% vs. 16.0% vs. 24.7%), highlighting the stronger discriminative ability of GRADE in assessing knowledge-intensive reasoning in image editing.

Moreover, closed-source models consistently outperform open-source ones, where the best-performing open-source model, Qwen-Edit-2511 [35] (2.7%), remains lower than all closed-source models, and most open-source models like OmniGen [38], Bagel [6], and FLUX.2 dev [16] only achieve near-zero or zero accuracy. As shown in Tab. 3, under the relaxed score, a similar performance gap between closed-source and open-source models is still observed, with most closed-source models scoring above 40% while open-source models largely remain below this level.

**Table 2:** Performance of different models across disciplines.

| Model | Phy | Sports | Chem | Math | Music | Econ | Hist | Geo | Bio | Comp |
|---|---|---|---|---|---|---|---|---|---|---|
| ▼ **Closed Source Models** | | | | | | | | | | |
| Nano Banana Pro [10] | **53.1** | **36.4** | 42.7 | **37.5** | **54.2** | 61.7 | **29.6** | **37.5** | **55.6** | **57.1** |
| Nano Banana 2 [11] | 35.9 | 31.8 | **44.0** | 33.9 | 37.5 | **71.7** | 22.2 | 21.9 | 38.9 | 42.9 |
| Seedream 5.0 [28] | 25.0 | 18.2 | 36.0 | 18.8 | 20.8 | 20.0 | 3.7 | 15.6 | 45.3 | 32.1 |
| GPT-Image-1.5 | 15.6 | 13.6 | 24.0 | 9.8 | 4.2 | 20.0 | 0.0 | 25.0 | 22.2 | 17.9 |
| FlUX.2 Max [16] | 3.1 | 9.1 | 21.3 | 7.1 | 12.5 | 3.3 | 3.7 | 12.5 | 29.6 | 21.4 |
| Nano Banana [33] | 1.6 | 4.6 | 12.0 | 6.3 | 16.7 | 15.0 | 3.7 | 9.4 | 13.0 | 14.3 |
| Seedream 4.5 [27] | 7.8 | 9.1 | 10.7 | 6.3 | 4.2 | 6.7 | 0.0 | 3.1 | 11.3 | 0.0 |
| GPT-Image-1.0 [22] | 9.4 | 0.0 | 8.0 | 4.5 | 0.0 | 5.0 | 0.0 | 3.1 | 11.1 | 14.3 |
| Flux.2 Pro [16] | 4.7 | 4.6 | 6.7 | 4.5 | 0.0 | 1.7 | 0.0 | 6.3 | 5.6 | 7.1 |
| Seedream 4.0 [29] | 0.0 | 2.3 | 5.3 | 4.5 | 4.2 | 3.3 | 0.0 | 3.1 | 3.8 | 0.0 |
| ▼ **Open Source Models** | | | | | | | | | | |
| Qwen-Edit-2511 [35] | 0.0 | 0.0 | 0.0 | 9.8 | 0.0 | 0.0 | 0.0 | 6.3 | 0.0 | 3.6 |
| Step-1X [19] (think+reflect) | 0.0 | 0.0 | 1.3 | 1.8 | 8.3 | 1.7 | 7.1 | 0.0 | 7.4 | 0.0 |
| FLUX.2 dev [16] | 0.0 | 2.3 | 1.3 | 3.6 | 4.2 | 0.0 | 0.0 | 3.1 | 5.6 | 0.0 |
| Step-1X [19] (think) | 0.0 | 0.0 | 2.7 | 0.0 | 4.2 | 0.0 | 3.7 | 6.3 | 1.9 | 0.0 |
| DreamOmni [37] | 0.0 | 0.0 | 1.3 | 0.0 | 4.2 | 1.7 | 0.0 | 0.0 | 3.7 | 0.0 |
| Step-1X [19] | 1.6 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 3.7 | 6.3 | 0.0 | 0.0 |
| Bagel | 0.0 | 0.0 | 1.3 | 0.0 | 4.2 | 0.0 | 0.0 | 0.0 | 1.9 | 0.0 |
| Bagel [6] (think) | 0.0 | 0.0 | 0.0 | 0.9 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| ICEdit [45] | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.7 | 0.0 | 0.0 | 0.0 | 0.0 |
| OmniGen [38] | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

From a dimension-wise perspective, the results reveal clear and consistent performance disparities across models. Closed-source models substantially outperform open-source ones: Nano Banana Pro achieves 77.5% in Reasoning, markedly higher than Seedream 5.0 (64.1%) and GPT-Image-1.5 (54.5%), while the best open-source model, Qwen-Edit-2511, reaches only 18.6%. Notably, even among closed-source models, reasoning ability varies significantly, with Nano Banana Pro exceeding Seedream 5.0 by 13.4%, indicating that discipline-informed editing exposes fine-grained differences in academic reasoning capability. In contrast, Visual Consistency and Logical Readability function as constraint-oriented criteria rather than ranking-oriented metrics. While many models achieve moderate scores on these dimensions, they play a critical diagnostic role by exposing distinct failure modes. For example, FLUX.2 dev exhibits a notably low consistency score (17.6%), indicating severe violations of instruction-scoped editing constraints and unintended structural modifications. Conversely, although DreamOmni [37] attains relatively high Visual Consistency scores (83.2%) and Logical Readability scores (89.1%), its overall performance remains low, largely because the model favors minimal or no changes to preserve the original image. These two dimensions complement Discipline Reasoning by enforcing structural and representational constraints, ensuring that high reasoning scores reflect valid and interpretable edits rather than superficial solutions.

As shown in Tab. 2, across disciplines, clear performance disparities emerge. Nano Banana Pro consistently achieves the highest scores across nearly all disciplines, indicating robust and stable discipline-informed reasoning rather than iso-

**Table 3:** Relaxed Score of different models across subjects.

| Model | Phy | Sports | Chem | Math | Music | Econ | Hist | Geo | Bio | Comp | Relaxed Score |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ▼ **Closed Source Models** | | | | | | | | | | | |
| Nano Banana Pro [10] | **84.0** | **75.7** | **85.6** | **75.9** | **82.8** | 89.5 | **81.7** | **80.5** | **89.0** | **90.1** | **82.9** |
| Nano Banana 2 [11] | 77.2 | 68.8 | 83.2 | 74.7 | 76.7 | **92.6** | 75.5 | 76.2 | 83.2 | 78.1 | 79.1 |
| Seedream 5.0 [28] | 74.1 | 72.9 | 79.7 | 67.4 | 62.6 | 79.5 | 66.2 | 71.5 | 82.6 | 74.8 | 73.7 |
| GPT-Image-1.5 [23] | 60.8 | 67.8 | 72.9 | 58.5 | 62.3 | 72.1 | 53.8 | 77.1 | 74.2 | 69.0 | 66.5 |
| Nano Banana [33] | 56.9 | 58.1 | 58.7 | 49.3 | 44.5 | 59.5 | 54.3 | 56.7 | 63.4 | 60.5 | 56.0 |
| FLUX.2 Max [16] | 54.7 | 56.0 | 63.2 | 45.4 | 38.6 | 53.0 | 64.2 | 60.6 | 69.7 | 57.8 | 55.7 |
| GPT-Image-1.0 [22] | 55.4 | 56.7 | 61.9 | 43.3 | 51.5 | 57.8 | 44.5 | 58.4 | 60.2 | 58.0 | 54.2 |
| Seedream 4.5 [27] | 53.6 | 54.4 | 57.4 | 40.5 | 40.7 | 55.8 | 35.6 | 43.2 | 60.2 | 44.9 | 49.7 |
| FLUX.2 Pro | 46.6 | 54.0 | 50.6 | 43.3 | 28.0 | 36.0 | 43.7 | 55.2 | 59.3 | 53.7 | 47.0 |
| Seedream 4.0 [29] | 41.8 | 51.8 | 48.1 | 36.0 | 32.4 | 49.3 | 28.9 | 41.7 | 49.3 | 47.2 | 43.1 |
| ▼ **Open Source Models** | | | | | | | | | | | |
| DreamOmni [37] | 40.6 | 52.7 | 50.6 | 40.1 | 47.4 | 44.9 | 35.5 | 35.9 | 47.0 | 48.9 | 44.3 |
| FLUX.2 dev [16] | 34.3 | 54.3 | 34.0 | 37.0 | 34.8 | 32.8 | 40.4 | 42.9 | 37.1 | 36.7 | 37.6 |
| Step-1X [19] (t+r) | 31.1 | 45.3 | 33.6 | 30.6 | 32.1 | 29.2 | 54.3 | 44.8 | 43.1 | 25.6 | 35.4 |
| Step-1X [19] (think) | 30.9 | 45.0 | 34.0 | 29.3 | 31.7 | 27.5 | 49.8 | 48.7 | 38.1 | 23.7 | 34.3 |
| Bagel [6] | 27.7 | 44.3 | 35.4 | 34.3 | 34.4 | 27.2 | 38.5 | 35.2 | 34.0 | 30.7 | 33.7 |
| Step-1X [19] | 30.2 | 41.8 | 27.0 | 27.9 | 32.0 | 29.3 | 50.8 | 45.8 | 35.0 | 27.5 | 32.6 |
| Bagel [6] (think) | 28.2 | 43.5 | 32.2 | 29.9 | 35.4 | 28.7 | 41.7 | 25.1 | 34.8 | 37.4 | 32.6 |
| Qwen-edit-2511 [35] | 27.7 | 45.1 | 13.7 | 39.5 | 19.8 | 28.6 | 36.9 | 33.0 | 27.5 | 23.9 | 30.0 |
| ICEdit [45] | 18.9 | 37.1 | 18.6 | 19.5 | 30.6 | 22.9 | 12.3 | 24.9 | 19.6 | 16.3 | 21.5 |
| OmniGen [38] | 19.5 | 44.0 | 15.7 | 15.3 | 30.0 | 11.5 | 24.8 | 23.4 | 21.2 | 31.8 | 21.1 |

lated domain strengths. Specifically, STEM-oriented disciplines such as Physics, Biology, and Mathematics exhibit the strongest differentiation: for example, in Physics, Nano Banana Pro achieves 53.1%, notably higher than Seedream 5.0 (25.0%) and GPT-Image-1.5 (15.6%), with similar gaps observed in Biology (55.6% vs. 45.3% and 22.2%, respectively). In contrast, humanities-related disciplines, particularly History and Geography, remain challenging even for closed-source models, with accuracy dropping sharply (e.g., Nano Banana Pro attains only 29.6%). Meanwhile, open-source models largely fail across disciplines, with accuracy frequently at or near zero.

### 4.2 Human Alignment

We analyze the alignment between our automated evaluation pipeline and human judgment on 68 samples uniformly sampled across disciplines, using outputs from three representative models (Nano Banana Pro, GPT-Image-1.5, Qwen-Edit-2511), where results are averaged across models and examined separately for each evaluation dimension. Human experts are required to follow the evaluation protocol to rate the generated images of Discipline Reasoning, Visual Consistency and Logical Readability. We collect annotations from five human experts and use the averaged human scores to mitigate individual bias. The alignment between human judgments and automated scores is measured using both mean absolute error (MAE) and standard deviation (STD), all values are normalized to the [0, 1] range. As shown in the last row of Tab. 4, the MAE val-

ues across all three dimensions are around 10% while the STD remains relatively stable across dimensions, indicating consistent variance in scoring deviations.

**Table 4: Alignment between automated evaluation and human judgments.** We report mean absolute error (MAE) and standard deviation (STD) of three evaluation dimensions.

| Judge Model | Reasoning | | Consistency | | Readability | |
|---|---|---|---|---|---|---|
| | **MAE** (↓) | **STD** (↓) | **MAE** (↓) | **STD** (↓) | **MAE** (↓) | **STD** (↓) |
| Qwen3-VL-235B | 0.1798 | 0.3504 | 0.1231 | 0.3523 | 0.0838 | 0.2944 |
| GPT-5 | 0.1519 | 0.3227 | 0.1677 | 0.4386 | 0.1221 | 0.3638 |
| **Gemini-3-Flash** | **0.1194** | **0.2834** | **0.0954** | **0.3047** | **0.0838** | **0.2944** |

### 4.3 Ablation Study

**Judge Model**. To study the impact of the judging model, we compare several commonly used MLLMs for automated evaluation, including an open-source model (Qwen3-VL-235B-Instruct [2]) and a closed-source model (GPT-5 [30]), against our default judge, Gemini-3-Flash [9]. All judges are evaluated on the same 68-sample subset by measuring their alignment with human annotations in Sec. 4.2. As shown in Tab. 4, Gemini-3-Flash consistently achieves the lowest MAE across all three evaluation dimensions, indicating the strongest alignment with human judgments. It also exhibits the lowest STD, reflecting more stable and consistent scoring behavior across samples. These results indicate that Gemini-3-Flash provides a reliable approximation of human evaluation and is well-suited for use as the judging model in our automated evaluation pipeline.

**Instruction Explicitness**. All instructions in our benchmark are designed to require implicit reasoning. For example, an instruction may ask for the product of a chemical reaction given a molecular structure, without explicitly describing how the structure should be transformed. To study the role of instruction formulation, we convert all instructions into explicit versions that directly specify the required knowledge-driven edits, while keeping the input and ground truth images unchanged. We select one open-source model (Qwen-Edit-2511) and one closed-source model (Nano Banana 2) with moderate overall performance on the full benchmark for this ablation study. Their performance under implicit and explicit instructions on the 68-sample subset is reported in Tab. 5. As shown, making the instructions explicit consistently improves performance for both models. The largest gains among the three metrics are observed in Discipline Reasoning, where Nano Banana 2 improves from 67.9% to 89.7% and Qwen-Edit-2511 from 18.9% to 44.7%. Overall, the open-source model Qwen-Edit-2511 exhibits a larger relative improvement, with accuracy increasing from 1.5% to 8.8%, suggesting that open-source models rely more on explicit guidance than closed-source models, reflecting a larger gap in implicit reasoning capability.

**Table 5:** Ablation study on instruction explicitness.

| Model | Reasoning | Consistency | Readability | Accuracy |
|---|---|---|---|---|
| Nano Banana 2 | 67.9 | 83.8 | 95.6 | 35.3 |
| w/ explicit inst. | 89.7 | 93.0 | 95.5 | 65.7 |
| Qwen-Edit-2511 | 18.9 | 40.0 | 47.8 | 1.5 |
| w/ explicit inst. | 44.7 | 60.0 | 81.6 | 8.8 |

### 4.4 Case Study

As shown in Fig. 4, we present a qualitative case study of six representative models that achieve relatively strong overall performance on the benchmark, including three closed-source (Nano Banana Pro, GPT-Image-1.5 and Seedream 5.0) and three open-source models (Qwen-Edit-2511, Step-1X (think+reflect) and DreamOmni).

For the example above, the diagram suggests that both Nano Banana Pro and GPT-Image-1.5 demonstrate an understanding of the concepts of the generatrix and the axis of rotation. Based on these cues, they further infer that the resulting shape should exhibit the characteristic geometry of being narrow at the top and bottom while wider in the middle. However, neither model is able to precisely reason about the spatial relationship between the generatrix and the rotation axis, and thus fails to conclude that the correct result should be an hourglass-shaped solid composed of two standard cones. Meanwhile, Seedream 5.0 correctly recognizes the concept of a solid of revolution but incorrectly infers the resulting shape to be a single cone. The remaining three open-source models fail to demonstrate an understanding of solids generated by rotation in three-dimensional space, indicating a clear limitation in spatial reasoning related to rotational geometry.

The other example further challenges the model's integrated reasoning capabilities. The model must first identify the missing text region, infer the underlying historical event, and then insert the corresponding textual content. Interestingly, we observe an unexpected phenomenon: none of the strong closed-source models correctly identified the missing position. Instead, they tended to insert additional information into other available empty spaces in the figure. In contrast, some of the open-source models (e.g., Qwen-Edit-2511 and Step-1X (think+reflect)) were able to correctly infer where the missing text should be placed, yet failed at the knowledge reasoning stage, leading to the incorrect identification of the corresponding year and historical event.

### 4.5 Error Analysis

Testing models' knowledge-based reasoning is the core of our benchmark. As illustrated in Fig. 5, we conduct a targeted error analysis of Nano Banana Pro to examine why current SoTA models still fall short, jointly analyzing its final outputs and chain-of-thought process. This enables us to pinpoint systematic

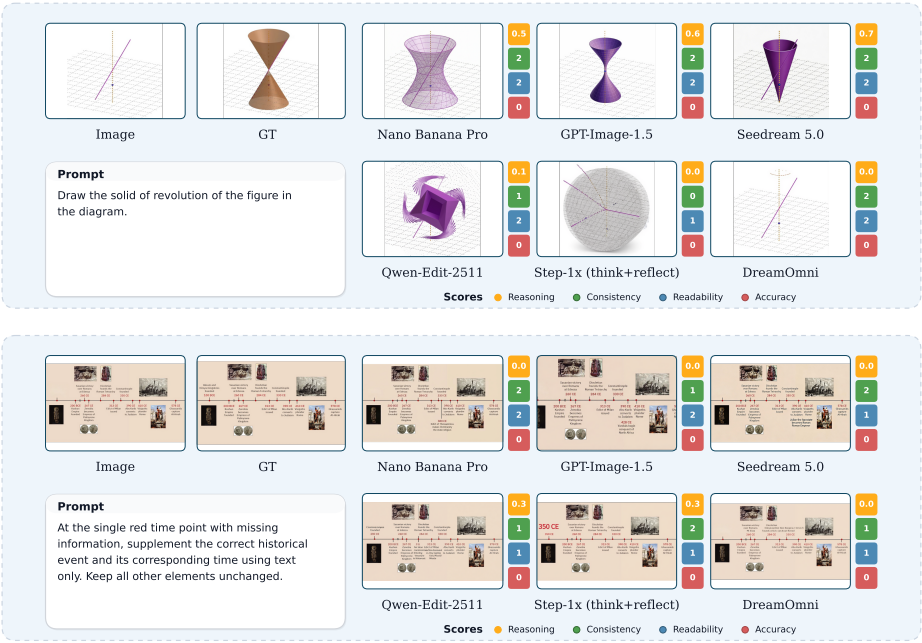


**Fig. 4:** Qualitative comparison of six representative high-performing models.

bottlenecks across perception, knowledge grounding, procedural execution, and synthesis, revealing whether failures stem from input misperception, inadequate knowledge retrieval, breakdowns in multi-step reasoning, or constraint violations during generation. We categorize them into four representative error types:

**(1) Image Recognition Error.** The model fails to robustly extract what each symbol is, where it is, and how elements align/connect in dense, structured visuals, so downstream reasoning is grounded on a wrong perceptual parse.

**(2) Knowledge Error.** When an image follows a discipline-specific canonical template (e.g., orbital energy diagrams, reaction-mechanism arrows, chart axes semantics, anatomical labels, geometric construction rules), failing to trigger the relevant priors makes the model treat the task as generic "shape completion/node adding," producing elements that are semantically invalid for the system.

**(3) Reasoning Process Error.** The model correctly recognizes entities and identifies the right methodology (e.g., shortest paths, conservation laws, grammar rules, geometric theorems), yet deviates during multi-step execution by missing constraints, state updates, or selection criteria, treating intermediate artifacts as final results.

**(4) Generation Process Error.** The model can correctly plan the target and constraints ("what to keep" and "which attributes must remain fixed"), but in final image synthesis, these are not reliably enforced at low-level control, so the output drifts due to priors, global style consistency, or pattern copying.

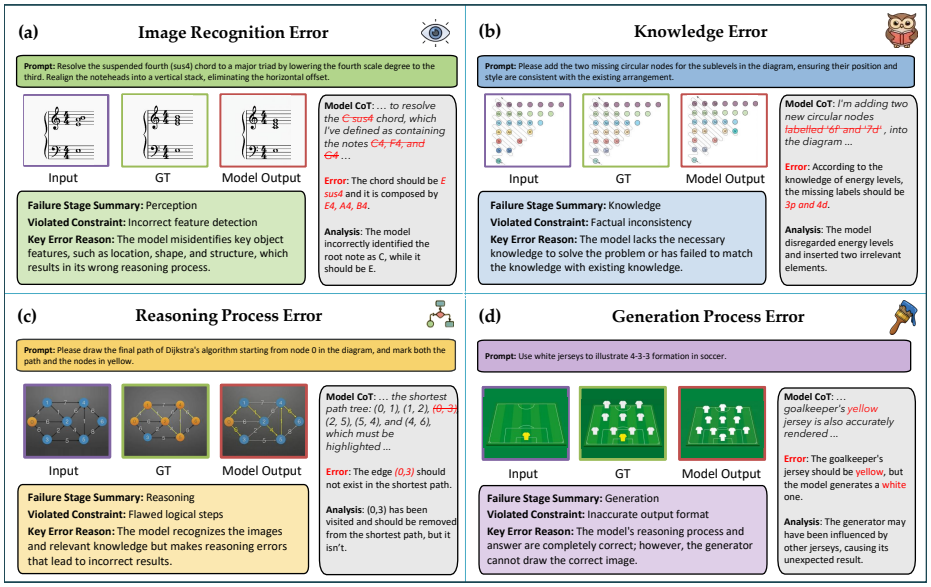


**Fig. 5: Error analysis of Nano Banana Pro.** GT: Ground truth image. Error types include: (a) Image recognition error: mis-parsed structural cues. (b) Knowledge error: failure to activate domain-specific priors produces semantically invalid elements. (c) Reasoning process error: correct methodology but flawed multi-step execution violates constraints. (d) Generation process error: correct planning but failure to enforce hard constraints during synthesis.

## 5 Conclusion

In this work, we present GRADE, a discipline-informed image editing benchmark that systematically evaluates editing performance grounded in discipline-informed knowledge across ten academic disciplines. GRADE comprises 520 carefully constructed samples of 10 disciplines and provides a multi-dimensional evaluation protocol that demonstrates strong alignment with human judgments. Through extensive experiments on a wide range of models, our benchmark reveals substantial performance gaps, particularly in applying discipline-informed reasoning under implicit instructions. These results highlight persistent limitations of current models in handling structured academic knowledge beyond visual realism, and point to the need for future unified multimodal models to better integrate discipline-informed knowledge, reasoning, and editing.

## Acknowledgement

This work was supported by the National Natural Science Foundation of China (62506229), Natural Science Foundation of Shanghai under 25ZR1402268, and Shanghai QiYuan Innovation Foundation.

## References


1. Analysis, A.: Image editing leaderboard. https://artificialanalysis.ai/image/leaderboard/editing/ (2026)
2. 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)
3. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Zhang, Q., Kreis, K., Aittala, M., Aila, T., Laine, S., et al.: ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022)
4. Chang, Y., Feng, Y., Sun, J., Ai, J., Li, C., Zhou, S.K., Zhang, K.: Sridbench: Benchmark of scientific research illustration drawing of image generation model. arXiv preprint arXiv:2505.22126 (2025)
5. Crowson, K., Biderman, S., Kornis, D., Stander, D., Hallahan, E., Castricato, L., Raff, E.: Vqgan-clip: Open domain image generation and editing with natural language guidance. In: European conference on computer vision. pp. 88–105. Springer (2022)
6. 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)
7. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
8. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment (2023), https://arxiv.org/abs/2310.11513
9. Google: Gemini 3 flash: frontier intelligence built for speed. https://blog.google/products-and-platforms/products/gemini/gemini-3-flash/ (2025)
10. Google: Introducing nano banana pro. https://blog.google/innovation-and-ai/products/nano-banana-pro/ (2025)
11. Google: Nano banana 2: Combining pro capabilities with lightning-fast speed. https://blog.google/innovation-and-ai/technology/ai/nano-banana-2/ (2025)
12. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium (2018), https://arxiv.org/abs/1706.08500
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
14. Hu, X., Wang, R., Fang, Y., Fu, B., Cheng, P., Yu, G.: Ella: Equip diffusion models with llm for enhanced semantic alignment (2024), https://arxiv.org/abs/2403.05135
15. Huang, Y., Xie, L., Wang, X., Yuan, Z., Cun, X., Ge, Y., Zhou, J., Dong, C., Huang, R., Zhang, R., Shan, Y.: Smartedit: Exploring complex instruction-based image editing with multimodal large language models (2023), https://arxiv.org/abs/2312.06739
16. Labs, B.F.: FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2 (2025)

17. Li, H., Tian, C., Shao, J., Zhu, X., Wang, Z., Zhu, J., Dou, W., Wang, X., Li, H., Lu, L., et al.: Synergen-vl: Towards synergistic image understanding and generation with vision experts and token folding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29767–29779 (2025)
18. Li, S., Huang, F., Zhang, L.: A survey of multimodal composite editing and retrieval (2024), https://arxiv.org/abs/2409.05405
19. Liu, S., Han, Y., Xing, P., Yin, F., Wang, R., Cheng, W., Liao, J., Wang, Y., Fu, H., Han, C., et al.: Step1x-edit: A practical framework for general image editing. arXiv preprint arXiv:2504.17761 (2025)
20. Luo, Y., Yuan, Y., Chen, J., Cai, H., Yue, Z., Yang, Y., Daha, F.Z., Li, J., Lian, Z.: Mmmg: A massive, multidisciplinary, multi-tier generation benchmark for text-to-image reasoning. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2025)
21. Niu, Y., Ning, M., Zheng, M., Jin, W., Lin, B., Jin, P., Liao, J., Feng, C., Ning, K., Zhu, B., Yuan, L.: Wise: A world knowledge-informed semantic evaluation for text-to-image generation (2025), https://arxiv.org/abs/2503.07265
22. OpenAI: Gpt-image-1. https://openai.com/index/image-generation-api/ (2025)
23. OpenAI: Gpt-image-1.5. https://openai.com/zh-Hans-CN/index/new-chatgpt-images-is-here/ (2025)
24. Phan, L., Gatti, A., Han, Z., Li, N., Hu, J., Zhang, H., Zhang, C.B.C., Shaaban, M., Ling, J., Shi, S., et al.: Humanity's last exam. arXiv preprint arXiv:2501.14249 (2025)
25. 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. Advances in neural information processing systems **35**, 36479–36494 (2022)
26. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans (2016), https://arxiv.org/abs/1606.03498
27. Seedream: Seedream 4.5. https://seed.bytedance.com/en/seedream4_5/ (2025)
28. Seedream: Seedream 5.0. https://seed.bytedance.com/en/seedream5_0_lite/ (2026)
29. Seedream, T., Chen, Y., Gao, Y., Gong, L., Guo, M., Guo, Q., Guo, Z., Hou, X., Huang, W., Huang, Y., et al.: Seedream 4.0: Toward next-generation multimodal image generation. arXiv preprint arXiv:2509.20427 (2025)
30. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)
31. Sun, K., Fang, R., Duan, C., Liu, X., Liu, X.: T2i-reasonbench: Benchmarking reasoning-informed text-to-image generation (2025), https://arxiv.org/abs/2508.17472
32. Team, C.: Chameleon: Mixed-modal early-fusion foundation models, 2024. URL https://arxiv. org/abs/2405.09818 **9**(8) (2024)
33. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
34. Wang, Z., Yin, P., Zhao, X., Tian, C., Qiao, Y., Wang, W., Dai, J., Luo, G.: Genexam: A multidisciplinary text-to-image exam. arXiv preprint arXiv:2509.14232 (2025)

35. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., ming Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., Zhang, H., Meng, H., Wei, H., Ni, J., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X., Wang, Y., Zhang, Y., Zhu, Y., Wu, Y., Cai, Y., Liu, Z.: Qwen-image technical report (2025), https://arxiv.org/abs/2508.02324
36. Wu, Y., Li, Z., Hu, X., Ye, X., Zeng, X., Yu, G., Zhu, W., Schiele, B., Yang, M.H., Yang, X.: Kris-bench: Benchmarking next-level intelligent image editing models. arXiv preprint arXiv:2505.16707 (2025)
37. Xia, B., Zhang, Y., Li, J., Wang, C., Wang, Y., Wu, X., Yu, B., Jia, J.: Dreamomni: Unified image generation and editing (2025), https://arxiv.org/abs/2412.17098
38. Xiao, S., Wang, Y., Zhou, J., Yuan, H., Xing, X., Yan, R., Li, C., Wang, S., Huang, T., Liu, Z.: Omnigen: Unified image generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13294–13304 (2025)
39. Ye, Y., He, X., Li, Z., Lin, B., Yuan, S., Yan, Z., Hou, B., Yuan, L.: Imgedit: A unified image editing dataset and benchmark. arXiv preprint arXiv:2505.20275 (2025)
40. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., Hutchinson, B., Han, W., Parekh, Z., Li, X., Zhang, H., Baldridge, J., Wu, Y.: Scaling autoregressive models for content-rich text-to-image generation (2022), https://arxiv.org/abs/2206.10789
41. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9556–9567 (2024)
42. Zhang, C., Wang, J., Wang, Y., Liang, Y., Yang, X., Li, Z., Huang, H., Li, X.: Unimodel: A visual-only framework for unified multimodal understanding and generation. arXiv preprint arXiv:2511.16917 (2025)
43. Zhang, D., Jiang, C., Xu, R., Chen, B., Jin, Z., Lu, Y., Zhang, J., Yong, L., Luo, J., Luo, S.: Worldgenbench: A world-knowledge-integrated benchmark for reasoning-driven text-to-image generation. arXiv preprint arXiv:2505.01490 (2025)
44. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
45. Zhang, Z., Xie, J., Lu, Y., Yang, Z., Yang, Y.: In-context edit: Enabling instructional image editing with in-context generation in large-scale diffusion transformers. In: Advances in Neural Information Processing Systems (NeurIPS) (2025), arXiv:2504.20690
46. Zhang, Z., Wang, J., Wen, F., Guo, Y., Zhao, X., Fang, X., Ding, S., Jia, Z., Xiao, J., Shen, Y., et al.: Large multimodal models evaluation: a survey. Science China Information Sciences **68**(12), 221301 (2025)
47. 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. arXiv preprint arXiv:2504.02826 (2025)