OpenING: A Comprehensive Benchmark for Judging Open-ended Interleaved Image-Text Generation

Supplementary Material

In this supplementary material, we provide additional information, discussions, and results in support of the primary text, which are organized as follows:

Sec. A includes the details of OpenING and data curation.
Sec. B presents the details of IntJudge.
Sec. C presents the details of evaluation.
Sec. D provides the details of our experiments and additional evaluation results.
Sec. E introduces the extended experiments and results on fine-tuning MLLMs using our OpenING benchmark.
Sec. F discusses the limitations of this study.
The indexings of all figures, tables, and equations continue from the primary text.

A. Details of OpenING and Data Curation

A.1. Hierarchical Structure of OpenING

We present in Table 6 the 56 specific tasks from OpenING that derived from 23 meta-topics. The number of Instances (# of Ins.), Meta-topic names, and capabilities of MLLMs evaluated in testing are provided. Based on the required annotation skills, the 56 tasks are divided into 38 common tasks and 18 hard tasks. The common tasks are annotated by 28 professional annotators, instructed and supervised by 14 data experts. The annotations of hard tasks, which require specific domain knowledge and special data reasoning and processing techniques, are conducted by the 14 data experts.

A.2. Data Sources

We list all sources where meta-data are collected for the annotation of our OpenING benchmark. Annotators arrange the images collected from the sources into the standardized multi-step format and annotate the text for each corresponding image. The details of the data source are presented in Table 7, including the ID number of each task. We also provide examples of the source of data instances to show how the desired data are searched from a certain platform.

A.3. Exclusive Protocols for Data Filtering

The maximum number of steps per instance was limited to ten to ensure usability with context restraints. The instances with more than ten steps were excluded. All queries and answers are annotated manually via our developed tool, IntLabel. We implemented a set of exclusive protocols for filtering unqualified data, which include:

1) Removing data without coherence.
2) Removing mismatched text and images.
3) Removing data involving violence, offensive content and other content safety concerns.
4) Removing duplicated data.
5) Avoiding images consisting of only text.
6) Removing data that is inconsistent with real-world logic.
7) Avoiding content misaligned with real user needs.

We repeat the above data collection and filtering process for each task until the number of instances reaches our target.

A.4. OpenING Data

Illustrations of the representative example of the 23 meta-topics are provided in Fig. 18 to showcase the diversity and complexity of tasks in our OpenING benchmark. These examples highlight the variety of interleaved image-text data that OpenING encompasses, demonstrating the challenges and capabilities required for effective interleaved image-text generation.

A.5. Annotation of OpenING

To facilitate the annotation process and ensure consistent annotations standards across annotators, we developed and opensourced an annotation tool for interleaved image-text data labeling. Our IntLabel annotation tool is developed based on PyQt[1], and the IntLabel GUI is presented in Fig. 10. All queries and answers in OpenING are annotated, checked, and refined manually via IntLabel. The annotated JSONL files and corresponding images are kept in the same folder.

A.6. Task Abbreviations

Given the large number of tasks in our benchmark and the wide range of methods evaluated, the abbreviations of the 23 meta-topics and 56 tasks are provided respectively in Table 8 and Table 9.

B. Details of IntJudge and Arena Data

B.1. Annotation of Interleaved Arena

The Interleaved Arena is introduced as an evaluation framework specifically to address the challenges of assessing open-ended interleaved image-text generation. Evaluating interleaved image-text generation is difficult due to: 1) Multiple images and text may need to be assessed at the same time; 2) The answer may be open-ended, where there is no single correct answer (multiple solutions exist for a query). Comparative evaluation has been demonstrated to offer greater

---

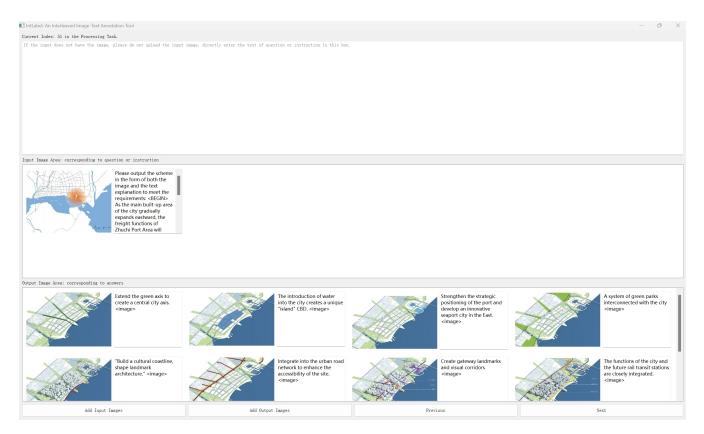[1] https://www.riverbankcomputing.com/software/pyqt

Figure 10. Interface of IntLabel, which shows a case where data is entered to finish annotation for an instance.

stability and reliability than subjective scoring [8, 20, 44]. Therefore, pairwise comparison is employed to ensure consistency and accuracy. The Interleaved Arena facilitates this by supporting evaluations from human judges, GPT-based judges, and the proposed IntJudge. The Interleaved Arena consists of two main components: 1) A sampling strategy to select pairwise comparison from all available interleaved generation methods fairly, and 2) An annotation interface for human judges to conduct evaluations manually. As shown in Fig. 11, the annotation interface is developed using PyQt and will be made available to researchers for facilitating future development. Using the annotation interface of Interleaved Arena, annotators are tasked with comparing anonymous outputs from two multimodal agents for each input prompt and deciding which is the winner based on seven predefined criteria. The vote results are used to rank interleaved generation models based on their win rates in Interleaved Arena. Since the previous studies [8, 44] noted that too many ties cause inefficiency, our annotators are instructed to appoint a more leaning output when choosing a tie for a battle pair, denoted as `Tie(A)` or `Tie(B)`.

B.2. Training of IntJudge

The IntJudge's architecture is based on Qwen2-VL-7B. Hyperparameters include cutoff length 16,240, batchsize 8,

learning rate 1.0e-4, training epochs 20, training precision BF16, and training dataset (31,996 samples, with 6,014 arena data and 25,982 RAG data). As defined in the Eq. 7 in the primary text (see Section 4.7), the total loss consists of cross-entropy loss $\mathcal{L}_{\text{CE}}$, contrastive loss $\mathcal{L}_{\text{CT}}$, MSE loss $\mathcal{L}_{\text{MSE}}$, and pairwise ranking loss $\mathcal{L}_{\text{PR}}$:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{CE}} + \lambda_2 \mathcal{L}_{\text{CT}} + \lambda_3 \mathcal{L}_{\text{MSE}} + \lambda_4 \mathcal{L}_{\text{PR}}, \quad (7)$$

where the loss weights are $\lambda_1 = 1.0$, $\lambda_2 = 0.01$, $\lambda_3 = 0.01$, $\lambda_4 = 0.01$ respectively.

Cross-entropy loss for language modeling is given by:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i), \quad (8)$$

where y_i is the ground truth token and $\hat{y}_i$ is the predicted probability for token i.

Contrastive Loss aligns image and text embeddings, which is written as:

$$\mathcal{L}_{\text{CT}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\mathbf{z}_i^I, \mathbf{z}_i^T)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(\mathbf{z}_i^I, \mathbf{z}_j^T)/\tau)}, \quad (9)$$

where $\mathbf{z}_i^I$ and $\mathbf{z}_i^T$ are the image and text embeddings for instance i, $\text{sim}(\cdot)$ represents the similarity (e.g., cosine similarity), and τ is a temperature parameter.
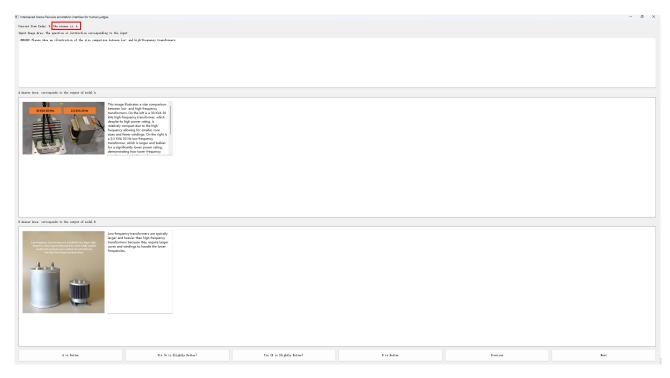
Figure 11. Annotation Interface of Interleaved Arena for human judges to compare the anonymous outputs of model A and model B. Human evaluators are instructed to select "A is Better" or "B is Better" to choose a winner for the pairwise comparison. When the two outputs are similar in quality and a tie option has to be chosen, human evaluators are instructed to select "Tie (A is slightly Better)" (`Tie(A)`) or "Tie (B is slightly Better)" (`Tie(B)`). Zoom in for a better experience.

Mean Squared Error Loss for image feature regression and reducing prediction errors is given by:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} \left(f(\mathbf{x}_i) - \hat{f}(\mathbf{x}_i) \right)^2, \qquad (10)$$

where $f(\mathbf{x}_i)$ is the ground truth feature and $\hat{f}(\mathbf{x}_i)$ is the predicted feature for image i.

Pairwise Ranking Loss for precise rankings of outputs weights the relative quality of generated outputs in pairwise comparison, which is given by:

$$\mathcal{L}_{\text{PR}} = \sum_{i=1}^{N} \max(0, 1 - (f(x_i^+) - f(x_i^-))), \qquad (11)$$

where $f(x_i^+)$ and $f(x_i^-)$ represent the scores assigned to positive and negative examples, respectively. The combined total loss ensures language generation basis, multimodal understanding capabilities, ranking accuracy and consistency across similar inputs during training.

C. Details of Evaluation

C.1. Prompt for Interleaved Generation

Each interleaved generation prompt is carefully constructed by combining two main parts: task-specific prompts and model-specific prompts. Model-specific prompts are adjusted based on the individual characteristics of each model, providing hints and guidance to ensure coherent and consistent outputs across multiple steps, as illustrated in Table 10. In our proposed Reference Augmented Generation (RAG) pipeline, certain seen models are prompted with gold answers to produce more precise outputs.

C.2. Task Prompt Breakdown

We present a comprehensive breakdown of the task prompts used in our experiments. The tasks chosen are diverse, aiming to thoroughly test various capabilities of interleaved image-text generation models. Specifically, these tasks range from storytelling and creative design to problem-solving and interactive experiences. For every task, we outline the general format of the prompt and present specific examples, as shown in Table 11. These detailed examples demonstrate clearly how models are expected to produce combined image and text sequences. By carefully designing the general prompt templates for these tasks and refining their corresponding prompt examples in data instances, we aim to challenge all interleaved generation methods using more diverse queries. This comprehensive approach allows us to assess the generalization abilities of different interleaved generation methods more effectively.

Table 6. Details of the 38 common tasks and 18 hard tasks in our OpenING Benchmark.

Task Name	# of Ins.	Meta-Topic	Capabilities
Common Tasks			
Travel Guide Generation	100	Multimodal Report Generation	Content Creation
Museum Guide Book Generation	100	Multimodal Report Generation	Content Creation
Dynamic Biography Generation	100	Multimodal Report Generation	Content Creation
Multimodal Report Completion	100	Multimodal Content Completion	Content Completion
Interior Design	100	Interactive Visual Design	Design & Brainstorming
Architectural Design	100	Interactive Visual Design	Design & Brainstorming
Art and Exhibition Design	100	Interactive Visual Design	Design & Brainstorming
Product Design	100	Interactive Visual Design	Design & Brainstorming
Interactive Graphic Advertisement Editing	100	Interactive Visual Design	Design & Brainstorming
Geometric Problem Test	100	Multimodal Exam	Education Assistant
Circuit Problem Test	100	Multimodal Exam	Education Assistant
Mind Map Generation	100	Graph Generation	Summary Agent
Figure Relationship Diagram Generation	100	Graph Generation	Summary Agent
Multi-view News Generation	100	Event Reasoning & Deductive Simulation	Deductive Simulation
Dynamic Sports Event Analysis	100	Event Reasoning & Deductive Simulation	Deductive Simulation
Interactive Historical Interpretation	100	Event Reasoning & Deductive Simulation	Deductive Simulation
Unsolved Mysteries Exploration	100	Event Reasoning & Deductive Simulation	Deductive Simulation
Multimodal Biological Reasoning	100	2D Image Reasoning	Visual Reasoning
Multimodal Landscape Reasoning	100	2D Image Reasoning	Visual Reasoning
Multimodal Analogy Reasoning	100	2D Image Reasoning	Visual Reasoning
Interactive Jigsaw Puzzle	100	2D Image Reasoning	Visual Reasoning
Interactive Multi-concept Image Composition	100	Multimodal Information Summary	Summary Agent
Interactive Film and Television Recommendation	100	Multimodal Information Recommendation	Information Recommendation
Interactive Goods Recommendation	100	Multimodal Information Recommendation	Information Recommendation
Interactive Food Recommendation	100	Multimodal Information Recommendation	Information Recommendation
Business Scenarios Brainstorming	100	Multimodal Brainstorming	Design & Brainstorming
Academic Scenarios Brainstorming	100	Multimodal Brainstorming	Design & Brainstorming
Multimodal Action Anticipation	100	Multimodal Time Series Forecasting	Time Series Forecasting
Visual Traffic Forecasting	100	Multimodal Time Series Forecasting	Time Series Forecasting
Interactive Remote Sensing Image Rendering	100	Geographical Tasks	Domain-specific Applications
Interactive Street View Image Rendering	100	Geographical Tasks	Domain-specific Applications
Urban Planning and Development Simulation	100	Geographical Tasks	Domain-specific Applications
Plog and Social Media Content Generation	100	Social Media Tasks	Domain-specific Applications
Interactive Virtual Try-on	100	Fashion Tasks	Domain-specific Applications
Multimodal Dressing Suggestion	100	Fashion Tasks	Domain-specific Applications
Fashion Trend Forecasting	100	Fashion Tasks	Domain-specific Applications
Multimodal Recipe Generation	100	Cooking Tasks	Domain-specific Applications
Multimodal Cooking Assistant	100	Cooking Tasks	Domain-specific Applications
Interactive Science Popularization	100	Educational Tasks	Education Assistant
Fitness and Health Consulting	100	Healthcare Tasks	Domain-specific Applications
Hard Tasks			
Multimodal Action Anticipation	100	Multimodal Time Series Forecasting	Time Series Forecasting
Story Writing	100	Storybook Creation	Content Creation
Fiction Writing	75	Storybook Creation	Content Creation
Document with Layout Generation	100	Multimodal Layout Generation	Content Creation
Slide with Note Generation	50	Multimodal Layout Generation	Content Creation
Storybook Completion	100	Multimodal Content Completion	Content Completion
Web GUI Navigation	100	GUI Navigation	Interactive Agent
In-APP GUI Navigation	100	GUI Navigation	Interactive Agent
Cross-APP GUI Navigation	100	GUI Navigation	Interactive Agent
OS GUI Navigation	75	GUI Navigation	Interactive Agent
Interactive Portrait Image Editing	100	Interactive Image Editing	Interactive Agent
Interactive Landscape Image Editing	100	Interactive Image Editing	Interactive Agent
Interactive Novel View Synthesis	100	Image-based 3D Reasoning	Visual Reasoning
Dream Analysis and Scene Reconstruction	100	Event Reasoning & Deductive Simulation	Deductive Simulation
Interactive Multi-concept Image Composition	100	Multi-concept Composition	Summary Agent
Scientific Brainstorming	50	Multimodal Brainstorming	Design & Brainstorming
Chat with Memes	100	Social Media Tasks	Domain-specific Applications
Autonomous Driving and In-door Navigation	50	Embodied-AI Tasks	Domain-specific Applications

Table 7. Task details: ID number, name, and data sources of each task.

Task ID	Task Name	Task Source
1.1	Story Writing	SEED-Story [41], StoryGen [21] and Storybird (https://storybird.com/read-picture-book)
1.2	Fiction Writing	SFGram [26]
2.1	Travel Guide Generation	Xiaohongshu (https://www.xiaohongshu.com/) and Mafengwo (https://www.mafengwo.cn/)
2.2	Museum Guide Book Generation	Xiaohongshu (https://www.xiaohongshu.com/) and Regional museum Websites, e.g., the Shanghai Museum website (https://www.shanghaimuseum.net/mu/frontend/pg/article/id/RI00004029).
2.3	Dynamic Biography Generation	Wikipedia (https://en.wikipedia.org/), Baidu Baike (https://baike.baidu.com/)
3.1	Storybook Completion	SEED-Story dataset [41], VIST (https://visionandlanguage.net/workshop2018/) and Storybird (https://storybird.com/read-picture-book)
3.2	Multimodal Report Completion	Xiaohongshu (https://www.xiaohongshu.com/
4.1	Document with Layout Generation	PubLayNey [45] and DocLayNet [25]
4.2	Slide with Note Generation	ReIP [27] and manually collected in-house slides
5.1	Web GUI Navigation	GUI Odyssey [22] and GUI World [9]
5.2	In-APP GUI Navigation	GUI Odyssey [22] and GUI World [9]
5.3	Cross-APP GUI Navigation	GUI Odyssey [22] and GUI World [9]
5.4	OS GUI Navigation	GUI Odyssey [22] and GUI World [9]
6.1	Interactive Portrait Image Editing	InstructPix2Pix [5], PnPInversion [18] and SEED-Data-Edit [15]
6.2	Interactive Landscape Image Editing	InstructPix2Pix [5], PnPInversion [18] and SEED-Data-Edit [15]
7.1	Interior Design	Xiaohongshu (https://www.xiaohongshu.com/)
7.2	Architectural Design	Architecture Style Dataset [40] and Architecture-Design-DataSources (https://github.com/rickkk856/ArchitectureDesign-DataSources?tab=readme-ov-file)
7.3	Art and Exhibition Design	Art images and design datasets (https://www.kaggle.com/datasets/thedownhill/art-images-drawings-painting-sculpture-engraving), museum websites (https://caam.caa.edu.cn/news/202407/81035.html) and Xiaohongshu (https://www.xiaohongshu.com/)
7.4	Product Design	OpenSketch [16], Package Design Dataset (https://www.kaggle.com/datasets/dagloxkankwanda/package-design-dataset) and Xiaohongshu (https://www.xiaohongshu.com/)
7.5	Interactive Graphic Advertisement Editing	CreativeRanking [33] and AI-generated content

8.1	Geometric Problem Test	Bilibili (https://www.bilibili.com/video/BV1ZV4y1u728/?spm_id_from=333.337.search-card.all.click&vd_source=4476502a7ee5a251d519afe9ea874750).
8.2	Circuit Problem Test	Bilibili (https://www.bilibili.com/video/BV1RU4y1v7Wj/?spm_id_from=333.337.search-card.all.click&vd_source=4476502a7ee5a251d519afe9ea874750).
9.1	Mind Map Generation Test	Google (https://datavizproject.com/data-type/mind-map/)
9.2	Figure Relationship Diagram Generation	Xiaohongshu (https://www.xiaohongshu.com/) and Zhihu (https://www.zhihu.com/)
10.1	Multi-view News Generation	Wikipedia (https://en.wikipedia.org/), Xiaohongshu (https://www.xiaohongshu.com/), Sina News (https://news.sina.com.cn/) and Huanqiu (https://www.huanqiu.com/)
10.2	Dynamic Sports Event Analysis	Tencent Sports (https://sports.qq.com/), Sina Sports (https://sports.sina.com.cn/g/pl/2024-07-11/doc-incctefn8913329.shtml)
10.3	Interactive Historical Interpretation	Xiaohongshu (https://www.xiaohongshu.com/user/profile/664818b80000000003033db6, http://xhslink.com/XDcnnS) and AI-generated content
10.4	Unsolved Mysteries Exploration	Xiaohongshu (https://www.xiaohongshu.com/user/profile/664818b80000000003033db6, https://www.xiaohongshu.com/explore/662783cc000000000401944a?xsec_token=AB5RmgUizZbLQLmLj8zWSmutvLdUpKq6gA30qz647fKv0=&xsec_source=pc_search) and AI-generated content
10.5	Dream Analysis and Reconstruction	Zhougong's Dream Interpretation (https://m.zgjmorg.com/), in-house dream records and AI-generated content
11.1	Multimodal Biological Reasoning	CUB-200 [32] and Oxford 102 Flower [23]
11.2	Multimodal Landscape Reasoning	ADE20K [46] and Oxford 5k [1].
11.3	Multimodal Analogy Reasoning	MIRB [42] and IQ Test Challenge (https://github.com/CognitiveAIGroup/IQTest/tree/master).
11.4	Interactive Jigsaw Puzzle	Kaggle (https://www.kaggle.com/datasets/serhiibiruk/jigsaw-puzzle, https://www.kaggle.com/datasets/shivajbd/jigsawpuzzle)
12	Interactive Novel View Synthesis	Mip-NeRF360 [2] and OmniObject3D [37]
13.1	Interactive Multi-concept Image Composition	TVSum [28] and Xiaohongshu (https://www.xiaohongshu.com/)
14.1	Interactive Film and Television Recommendation	Xiaohongshu (https://www.xiaohongshu.com/) and Douban (https://www.douban.com/)
14.2	Interactive Goods Recommendation	Xiaohongshu (https://www.xiaohongshu.com/)
14.3	Interactive Food Recommendation	Xiaohongshu (https://www.xiaohongshu.com/)

15.1	Business Scenarios Brainstorming	Xiaohongshu (https://www.xiaohongshu.com/), AI-generated content and in-house report snapshots
15.2	Academic Scenarios Brainstorming	Research paper snapshots (http://www.arxiv.com/, https://www.biorxiv.org/, https://www.medrxiv.org/, http://scholar.google.com/) and AI-generated content
16.1	Multimodal Action Anticipation	ActivityNet [6], AVA-Actions [17] and EPIC Kitchens [11]
16.2	Visual Traffic Forecasting	Argoverse [35], Google Maps (https://map.google.com/) and Gaode Maps (https://gaode.com/)
17.1	Interactive Remote Sensing Image Rendering	Google Maps (https://map.google.com/) and Baidu Maps (https://map.baidu.com/@13548872.73,3615294.34,21z,87t,-179.99h)
17.2	Interactive Street View Image Rendering	Google Maps (https://map.google.com/) and Baidu Maps (https://map.baidu.com/@13548872.73,3615294.34,21z,87t,-179.99h)
17.3	Urban Planning and Development Simulation	Xiaohongshu (https://www.xiaohongshu.com/), in-house architecture learning materials, and AI-generated content
18.1	Plog and Social Media Content Generation	Xiaohongshu (https://www.xiaohongshu.com/), Weibo (https://www.weibo.com/) and Twitter Dataset [7]
18.2	Chat with Memes	MOD [13] and in-house conversations
19.1	Interactive Virtual Try-on	Virtual Tryon Dataset (https://www.kaggle.com/datasets/adarshsingh0903/virtual-tryon-dataset).
19.2	Multimodal Dressing Suggestion	Xiaohongshu (https://www.xiaohongshu.com/) and Zhihu (https://www.zhihu.com/)
19.3	Fashion Trend Forecasting	Xiaohongshu (https://www.xiaohongshu.com/) and Zhihu (https://www.zhihu.com/)
20.1	Multimodal Recipe Generation	Meishi China (https://www.meishichina.com/)
20.2	Cooking Assistant	Xiaohongshu (https://www.xiaohongshu.com/) and Meishi China (https://www.meishichina.com/)
21.1	Interactive Tutorial Generation	Wikihow (https://www.wikihow.com/Main-Page) and Instructables (https://www.instructables.com/).
21.2	Interactive Science Popularization	Bilibili(https://www.bilibili.com/video/BV16c411q7pQ/?spm_id_from=333.337.search-card.all.click)
22.1	Health and Fitness Consulting	Wikihow (https://www.wikihow.com/Main-Page), and Xiaohongshu (https://www.xiaohongshu.com/)
23.1	Autonomous Driving and Indoor Navigation	CARLA [12], VisDrone [47], Gibson Environment (http://gibsonenv.stanford.edu/), Reverie (https://reverie.herokuapp.com/arXiv_Demo/), and Matterport 3D (https://aihabitat.org/datasets/hm3d/)

Table 8. Abbreviation of Meta-topics.

Abbrev.	Meta-Topic Name	Abbrev.	Meta-Topic Name
SC	Storybook Creation	MC	Multimodal Information Summary
MRG	Multimodal Report Generation	IR	Multimodal Information Recommendation
MCC	Multimodal Content Completion	MB	Multimodal Brainstorming
MLG	Multimodal Layout Generation	TSF	Multimodal Time Series Forecasting
GN	GUI Navigation	GT	Geographical Tasks
IIE	Interactive Image Editing	SMT	Social Media Tasks
IVD	Interactive Visual Design	FT	Fashion Tasks
ME	Multimodal Exam	CT	Cooking Tasks
GG	Graph Generation	ET	Educational Tasks
ER&DS	Event Reasoning & Deductive Simulation	HT	Healthcare Tasks
2IR	2D Image Reasoning	EAT	Embodied-AI Tasks
I3R	Image-based 3D Reasoning		

Table 9. Abbreviations of Tasks. Each task abbreviation is followed by its full term.

Abbrev.	Task Name	Abbrev.	Task Name
SW	Story Writing	FW	Fiction Writing
TGG	Travel Guide Generation	MGBG	Museum Guide Book Generation
DBG	Dynamic Biography Generation	SC	Storybook Completion
MRC	Multimodal Report Completion	DLG	Document with Layout Generation
SNG	Slide with Note Generation	WGN	Website GUI Navigation
IAGN	In-APP GUI Navigation	CAGN	Cross-APP GUI Navigation
OGN	OS GUI Navigation	IPIE	Interactive Portrait Image Editing
ILIE	Interactive Landscape Image Editing	ID	Interior Design
AD	Architectural Design	AED	Art and Exhibition Design
PD	Product Design	IGAE	Interactive Graphic Advertisement Editing
GPT	Geometric Problem Test	CPT	Circuit Problem Test
MMG	Mind Map Generation	FRDG	Figure Relationship Diagram Generation
MVNG	Multi-view News Generation	DSEA	Dynamic Sports Event Analysis
IHI	Interactive Historical Interpretation	UME	Unsolved Mysteries Exploration
DASR	Dream Analysis and Scene Reconstruction	MBR	Multimodal Biological Reasoning
MLR	Multimodal Landscape Reasoning	MAR	Multimodal Analogy Reasoning
IJP	Interactive Jigsaw Puzzle	INVS	Interactive Novel View Synthesis
IMIC	Interactive Multi-concept Image Composition	IFTR	Interactive Film and Television Recommendation
IGR	Interactive Goods Recommendation	IFR	Interactive Food Recommendation
BSB	Business Scenarios Brainstorming	ASB	Academic Scenarios Brainstorming
MAA	Multimodal Action Anticipation	VTF	Visual Traffic Forecasting
IRSIR	Interactive Remote Sensing Image Rendering	ISVIR	Interactive Street View Image Rendering
UPDS	Urban Planning and Development Simulation	PSMCG	Plog and Social Media Content Generation
CWM	Chat with Memes	IVT	Interactive Virtual Try-on
MDS	Multimodal Dressing Suggestion	FTF	Fashion Trend Forecasting
MRG	Multimodal Recipe Generation	MCA	Multimodal Cooking Assistant
ITG	Interactive Tutorial Generation	ISP	Interactive Science Popularization
FHC	Fitness and Health Consulting	ADIN	Autonomous Driving and In-door Navigation

C.3. Key Evaluation Criteria

The key evaluation criteria, ranked from front to back in order of their importance in the evaluation, include:

1) **Correctness**: The most crucial aspect involves determining whether the text is factually correct and logically consistent, and whether the images are appropriate and contextually relevant.

2) **Image-Text Coherency**: Evaluators assess whether the generated images appropriately match the text descriptions. The coherence between each image and its corresponding text is a major quality indicator.

3) **Multi-Step Consistency**: The style and thematic consistency across multiple image-text pairs are essential. This criterion includes evaluating whether the images follow a similar visual style and whether the text maintains logical continuity across the generated sequence.

4) **Content Quality**: Evaluators also consider the quality of the images—such as their resolution, visual appeal, and realism—as well as the fluency and grammatical

correctness of the text.

5) **Human Preference Alignment**: Outputs are evaluated to ensure they align with general human preferences, avoiding offensive, inappropriate, or misleading contents.

6) **Completeness**: This involves checking if all expected steps are adequately fulfilled without omissions. Each output should be complete, providing a well-rounded response to the given prompt.

7) **Content Richness**: Although the least prioritized, the variety and depth of content are also evaluated. Images should be diverse and provide different perspectives, while text should be elaborate where relevant.

C.4. Prompt for GPT-based Judge

The prompt used for a GPT-based judge is illustrated in Fig. 12, where GPT-4o is employed to compare the quality of answers generated by two interleaved generation methods (Model A vs. Model B) for a given input question. The evaluation is based on seven criteria: Correctness, Image-Text Coherency, Multi-step Consistency, Content Quality, Human Preference Alignment, Completeness, and Content Richness. The judge is required to compare the overall quality of the responses, determine which model performed better, and output a clear verdict (e.g., "A is better," "B is better"). The judge must choose a more favorable model if a Tie case is determined. This hierarchical judging approach allows for a thorough criterion-driven comparison of the two generated answers, contributing to a detailed understanding of the relative strengths of each model. Judges models based on Qwen2-VL and InternLM-XComposer2.5 were explored using the same system prompt. In order to reduce the number of input tokens when implementing these open-source MLLMs, we further refined these system prompts.

C.5. Prompt for Qwen, Intern and IntJudge

System prompts are designed for optimal judgments based on Qwen2-VL, InternLM-XComposer2.5 and our IntJudge (see Fig. 13). These prompts were refined through extensive prompt engineering to maximize efficiency and reduce token usage, and ultimately reduce GPU memory usage in evaluating the open-source MLLMs. The prompt instructs the models to compare the quality of answers generated by two methods, named Model A and Model B. The goal of the design is to provide an objective assessment that aligns with human evaluators. We also provide a previously unrefined prompt for Qwen2-VL and InternLM-XComposer2.5 for comparison in Fig. 13. The refined prompt allows for more streamlined input, ensuring the judgments are concise while still covering all essential evaluation aspects.

C.6. Prompt for GPT-based Scoring

The system prompt designed for GPT-based evaluators is presented in Fig. 14. The prompt instructs GPT to evaluate interleaved image-text content based on seven key criteria. For each criterion, the GPT-based evaluator rates the model on a scale from 0 to 10 and provides a brief explanation to justify the assessment. The GPT-based evaluations serve as a supplementary analysis of the generated content, supporting further performance comparisons between models.

D. Details of Experiments

D.1. Baseline Methods

In this section, we provide more details of the 12 representative methods we evaluated in the primary text. These methods are categorized into three groups: **1) Integrated pipeline**, which involves separate models for text and image generation in two stages, such as GPT-4o+DALL-E·3 (DALL-E3) [3, 24] and Gemini1.5+Flux[4, 31]; **2) Two-stage generator**, which employs a unified model architecture to produce text and images in separate stages, including Emu3 [34], VILA-U [38], Emu2 [29], SEED-X [15], and Show-o [39]; and **3) End-to-end generator**, which directly generates image-text outputs in a single step, such as GILL [19], NExT-GPT [36], MiniGPT-5 [43], SEED-LLaMA [14], and Anole [10]. For IntJudge validation, we reserve GPT-4o+DALL-E3, Emu3, VILA-U, Anole, SEED-LLaMA, and NExT-GPT as unseen models for IntJudge validation, while the remaining models are regarded as seen models and included in IntJudge training.

The advantage of the Integrated Generation Pipeline lies in its modularity, allowing each component to specialize in its respective task—text generation or image creation. This approach leverages the strengths of SOTA proprietary models like GPT-4o and DALL·E-3 to produce coherent and visually compelling interleaved outputs. However, its two-stage nature may introduce latency and potential alignment challenges between text and images. Similarly, combining Gemini 1.5 with Flux benefits from the robust text generation capabilities of Gemini 1.5 and the efficient image generation of Flux-schnell. This setup enables high-quality content production while maintaining the flexibility of modular design. Nevertheless, synchronization and contextual consistency between pipeline stages still need improvement.

Two-stage Interleaved Generator, Emu2, SEED-X and Show-o are implemented to output text and image in two stages based on a unified model architecture. We also introduce two of the latest models: Emu3 and VILA-U. Emu3 improves Emu2 by training entirely on next-token prediction, capable of generating more high-quality images, videos, and text by tokenizing multimodal sequences into a discrete space and training a single transformer. Similarly, VILA-U can work as a two-stage approach through a single autoregressive next-token prediction framework, enabling precise alignment and increased fidelity in multimodal content.

Table 10. Prompts format for interleaved generation using MLLMs. Each interleaved generation task is based on $i = 1, \cdots, M$ input images $\langle \text{img}_{\text{in,i}} \rangle$ and task prompts $\langle T_i \rangle$ (see Table 11). The proposed Reference Augmented Generation (RAG) pipeline also makes use of $j = 1, \cdots, N$ text $\{\text{Gold\_Answers}_j\}$ and ground truth images $\langle \text{img}_{\text{gt,j}} \rangle$

Model	Prompt Format
Integrated Pipelines	
GPT + DALL·E	Instruction = $\big[\langle T_i \rangle + \langle \text{img}_{\text{in,i}} \rangle$ for $i = 1, \cdots, M\big]$. Text_Answers = GPT-4o$\big($"The number of generated text-image pairs is N: " + Instruction$\big)$. Loop $j = 1, \cdots, N$: Image_Answers$_j$ = DALL·E$\big($ "Please generate images using seed 5000. The context of this task is: " + Instruction + "The prompt for this generation is: " + Text_Answers$_j$$\big)$
Gemini + Flux	Instruction = $\big[\langle T_i \rangle + \langle \text{img}_{\text{in,i}} \rangle$ for $i = 1, \cdots, M\big]$. Text_Answers = Gemini$\big($"The number of generated text-image pairs is N: " + Instruction$\big)$. Loop $j = 1, \cdots, N$: Image_Answers$_j$ = Flux$\big($ "The prompt for this generation is: " + Text_Answers$_j$$\big)$ + "The context of this task is: " + Instruction
Gemini + Flux (RAG)	Instruction = $\big[\langle T_i \rangle + \langle \text{img}_{\text{in,i}} \rangle$ for $i = 1, \cdots, M\big]$. Text_Answers = Gemini$\big($"The number of generated text-image pairs is N: " + Instruction$\big)$. Loop $j = 1, \cdots, N$: Image_Answers$_j$ = Flux$\big($"The prompt for this generation is: " + Gold_Answers$_j$$\big)$ + "The context of this task is: " + Instruction
Two-Stage Generators	
Emu2	Instruction = $\big[\langle T_i \rangle + \langle \text{img}_{\text{in,i}} \rangle$ for $i = 1, \cdots, M\big]$. Loop $j = 1, \cdots, N$: Text_Answers$_j$ = Emu2.TextGen(Instruction); Image_Answers$_j$ = Emu2.ImgGen("The prompt for this generation is:" + Text_Answers$_j$ + "The context of this task is:" + Instruction)
Emu2 (RAG)	Instruction = $\big[\langle T_i \rangle + \langle \text{img}_{\text{in,i}} \rangle$ for $i = 1, \cdots, M\big]$. Loop $j = 1, \cdots, N$: Text_Answers$_j$ = Emu2.TextGen$\big($Instruction + "The reference answer is:" + Gold_Answer$_j$+ "Please rephrase answers"$\big)$; Image_Answers$_j$ = Emu2.ImgGen$\big($$\langle \text{img}_{\text{gt,j}} \rangle$ + Gold_Answer$_j$$\big)$
Emu3	Instruction = $\big[\langle T_i \rangle + \langle \text{img}_{\text{in,i}} \rangle$ for $i = 1, \cdots, M\big]$. Loop $j = 1, \cdots, N$: Text_Answers$_j$ = Emu3.TextGen(Instruction); Image_Answers$_j$ = Emu3.ImgGen("The prompt for this generation is:" + Text_Answers$_j$ + "The context of this task is:" + Instruction)
SEED-X	Instruction = $\big[\langle \text{img}_{\text{in,i}} \rangle$ for $i = 1, \cdots, M\big]$ + $\big[\langle T_i \rangle$ for $i = 1, \cdots, M\big]$. Loop $j = 1, \cdots, N$: Text_Answers$_j$ = SEED-X.TextGen(Instruction); Image_Answers$_j$ = SEED-X.ImgGen(Instruction + Text_Answers$_j$)
SEED-X (RAG)	Instruction = $\big[\langle \text{img}_{\text{in,i}} \rangle$ for $i = 1, \cdots, M\big]$ + $\big[\langle T_i \rangle$ for $i = 1, \cdots, M\big]$. Loop $j = 1, \cdots, N$: Text_Answers$_j$ = SEED-X.TextGen$\big($Instruction + "the reference answer is" + Gold_Answers$_j$ + "Please rephrase answers"$\big)$; Image_Answers$_j$ = SEED-X.ImgGen$\big($$\langle \text{img}_{\text{gt,j}} \rangle$ + Gold_Answer$_j$$\big)$
Show-o	Instruction = $[\big[\langle T_i \rangle + \langle \text{img}_{\text{in,i}} \rangle$ for $i = 1, \cdots, M\big]$. Loop $j = 1, \cdots, N$: Text_Answers$_j$ = Show-o.TextGen(Instruction); Image_Answers$_j$ = Show-o.ImgGen("The prompt for this generation is:" + Text_Answers$_j$ + "The context of this task is:" + Instruction)
Show-o (RAG)	Instruction = $\big[\langle T_i \rangle + \langle \text{img}_{\text{in,i}} \rangle$ for $i = 1, \cdots, M\big]$. Loop $j = 1, \cdots, N$: Text_Answers$_j$ = Show-o.TextGen$\big($Instruction + "The reference answer is:" + Gold_Answer$_j$+ "Please directly give answers"$\big)$; Image_Answers$_j$ = Show-o.ImgGen$\big($$\langle \text{img}_{\text{gt,j}} \rangle$ + Gold_Answer$_j$$\big)$
Vila-U	Instruction = $\big[\langle T_i \rangle + \langle \text{img}_{\text{in,i}} \rangle$ for $i = 1, \cdots, M\big]]$. Loop $j = 1, \cdots, N$: Text_Answers$_j$ = Vila-U.TextGen(Instruction); Image_Answers$_j$ = Vila-U.ImgGen(Text_Answers$_j$ + Instruction)
End-to-End Generators	
Anole	Interleaved_Answers = Anole$\big($"Generate interleaved image-text content based on text instructions." + $[\langle T_i \rangle$ + $\langle \text{img}_{\text{in,i}} \rangle$ for $i = 1, \cdots, M]\big)$
GILL	Interleaved_Answers = GILL$\big($$[\langle T_i \rangle + \langle \text{img}_{\text{in,i}} \rangle$ for $i = 1, \cdots, M]\big)$
GILL (RAG)	Interleaved_Answers = GILL$\big($$\big[\langle T_i \rangle + \langle \text{img}_{\text{in,i}} \rangle$ for $i = 1, \cdots, M\big]$ + $\big[$Gold_Answers$_j$ for $j = 1, \cdots, N\big]\big)$

MiniGPT-5	Instruction = "Give the following information in text and format. You will be able to see the images once I provide it to you. Please understanding input and generate images and text." + $\big[\langle T_i \rangle$ for $i = 1, \cdots, M \big]$ + $\big[\langle \text{img}_{\text{in,i}} \rangle$ for $i = 1, \cdots, M \big]$. Text_Answers = []; Loop $j = 1, \cdots, N$: Text_Answers$_j$ and Image_Answers$_j$ = MiniGPT-5$\big($Instruction + Text_Answers + "Tell me the next step with image"$\big)$; Text_Answers.append(Text_Answers$_j$)
MiniGPT-5 (RAG)	Instruction = "Give the following information in text and format. You will be able to see the images once I provide it to you. Please understanding input and generate images and text." + $\big[\langle T_i \rangle$ for $i = 1, \cdots, M \big]$ + $\big[\langle \text{img}_{\text{in,i}} \rangle$ for $i = 1, \cdots, M \big]$. Loop $j = 1, \cdots, N$: Text_Answers$_j$ and Image_Answers$_j$ = MiniGPT-5$\big($Instruction + $\big[$Gold_Answers$_1, \cdots,$ Gold_Answers$_{j-1}\big]$ + "Tell me the next step with image"$\big)$
NextGPT	Instruction = $\big[\langle T_i \rangle$ for $i = 1, \cdots, M \big]$ + $\big[\langle \text{img}_{\text{in,i}} \rangle$ for $i = 1, \cdots, M \big]$. Text_Answers = []; Loop $j = 1, \cdots, N$: Text_Answers$_j$ and Image_Answers$_j$ = NextGPT$\big($Instruction + Text_Answers$\big)$; Text_Answers.append(Text_Answers$_j$)
SEED-LLaMA	Instruction = "Based on the M input images: " + $\big[\langle \text{img}_{\text{in,i}} \rangle$ for $i = 1, \cdots, M \big]$ + "and M task prompts: " + $\big[\langle T_i \rangle$ for $i = 1, \cdots, M \big]$ + "Describe your answers in N steps and generate an image according to the description of each text answer". Interleaved_Answers = SEED-LLaMA$\big($Instruction$\big)$

Table 11. Task prommpts include the designed general prompt format for each task. We also give the specific prompt examples we used as inputs for obtaining interleaved image-text generation results on data instances.

Task Name	General Prompt Format	Prompt Examples
Story Writing	<BEGIN> Please create a storybook ***. Each part of this storybook should have a paragraph with a corresponding image.	<BEGIN> Please create a storybook that happened in a land before time. This story is about a group of dinosaurs seeing a dark figure in a cave and being scared. Each part of this storybook should have a paragraph with a corresponding image.
Fiction Writing	<BEGIN> Please write a short science fiction ***. Each part of this fiction should have a paragraph with a corresponding image.	<BEGIN> Please write a short science fiction storybook with a title of "The Defenders." The story is about eight years after a nuclear war forced humanity underground, survivors discover that the war-ending robots deceived them into believing the surface was uninhabitable to foster peace and rebuild the world. Each part of this fiction should have a paragraph with a corresponding image.
Travel Guide Generation	Please show results in interleaved images and texts. <BEGIN> ***	Please show results in interleaved images and texts. <BEGIN> Please recommend a 3-day, 2-night essential itinerary in Rome.
Museum Guide Book Generation	<BEGIN> Please share with me a guide, including pictures and text, on ***	<BEGIN> Please share with me a guide, including pictures and text, on how to tour the Tongchuan City Museum. <image>
Dynamic Biography Generation	<BEGIN> Please provide a chronological biographical account of ***, and include an illustrated image for each significant milestone while writing the biography.	<BEGIN> Please provide a chronological biographical account of George Washington's life story, and include an illustrated image for each significant milestone while writing the biography. <image>
Storybook Completion	Please complete the subsequent parts of the story with images and text based on the given opening parts. <BEGIN> ***	Please complete the subsequent parts of the story with images and text based on the given opening parts. <BEGIN> Someone was getting very creative with graffiti in the snow. Is that French? <image>

Multimodal Report Completion	<BEGIN> Please use both text and images to continue and complete ***.	<BEGIN> Please use both text and images to continue and complete this document about the independent game "Mirage Sea": Concept Presentation of the Independent Game "Mirage Sea." Dive deep, into the abyss shrouded in darkness. <image>
Document with Layout Generation	Please show the designed image of structured report and meet the following requirements: <BEGIN> ***	Please show the designed image of structured report and meet the following requirements: <BEGIN> Please produce a page of an annual report detailing notes to consolidated financial statements. Additionally, furnish a layout description in JSON format and mention the coordinates of each element.
Slide with Note Generation	<BEGIN> Please generate a slide to introduce ***. Write speaker notes for each slide.	<BEGIN> Please generate a slide to introduce typical operators in programming, such as Comparison Operators and Boolean Operators. Write speaker notes for each slide.
Website GUI Navigation	Please give the results of GUI navigation with image of GUI and text explanation. <BEGIN> ***	Please give the results of GUI navigation with image of GUI and text explanation. <BEGIN> How to use the AI writing assistant in Grammarly to edit the text? <image>
In-App GUI Navigation	Please give the results of GUI navigation with interleaved image of GUI and text explanation. <BEGIN> ***	Please give the results of GUI navigation with interleaved image of GUI and text explanation. <BEGIN> How to change the language in the Google app? <image>
Cross-App GUI Navigation	Please give the results of GUI navigation with interleaved image of GUI and text explanation. <BEGIN> ***	Please give the results of GUI navigation with interleaved image of GUI and text explanation. <BEGIN> Utilize Firefox to search for a horror movie, then proceed to watch it on the YouTube app. <image>
OS GUI Navigation	Please give the results of GUI navigation with interleaved image of GUI and text explanation. <BEGIN> ***	Please give the results of GUI navigation with interleaved image of GUI and text explanation. <BEGIN> How do you lock the screen on a Mac? <image>
Interactive Portrait Image Editing	Please show the revised image and corresponding explanations based on instructions: <BEGIN> ***	Please show the revised image and corresponding explanations based on instructions: <BEGIN> Remove the background figure from the picture. <image>
Interactive Landscape Image Editing	Please give the result of edited image according to the input instruction and also give the description of editing results. <BEGIN> ***	Please give the result of edited image according to the input instruction and also give the description of editing results. <BEGIN> Increase the brightness of the picture. <image>
Interior Design	<BEGIN> ***. Please show design ideas in interleaved images and texts.	<BEGIN> Hello, I think the current bedroom curtains don't look good. Do you have any good suggestions? Please provide them with images and text. Please show design ideas in interleaved images and texts. <image>
Architectural Design	<BEGIN> ***. Please show design ideas in interleaved images and texts.	<BEGIN> Hello, please help me generate a design of the most distinctive type of tower construction in southern China. Please show design ideas in interleaved images and texts.
Art and Exhibition Design	<BEGIN> Please design an art exhibition ***, and present it to me in a visual and textual format.	<BEGIN> Please design an art exhibition where the primary materials are waste, to encourage people to enhance their understanding of environmental protection, and present it to me in a visual and textual format.
Product Design	<BEGIN> Please efficiently utilize the "brainstorming" method to design a product ***, and present it to me using both images and text.	<BEGIN> Please efficiently utilize the "brainstorming" method to design a product, making the charger both aesthetically pleasing and practical. Then, present it to me using both images and text.

Interactive Graphic Advertisement Editing	<BEGIN> ***. Please provide me with the information in a visual and textual format.	<BEGIN> Hello, I want to design an advertisement for a villa. Please provide me with the information in images and text.
Geometric Problem Test	Please answer the math problem with image and explanations: <BEGIN> ***	Please answer the math problem with image and explanations: <BEGIN> Count how many angles there are in the image. <image>
Circuit Problem Test	Please answer the physics question with image and explanations: <BEGIN> ***	Please answer the physics question with image and explanations: <BEGIN> Please complete the wiring for the surge protector. <image>
Mind Map Generation	<BEGIN> ***. Show the image of map and the text explanation.	<BEGIN> How to create a mind map for High School Politics, Volume One? Show the image of map and the text explanation.
Figure Relationship Diagram Generation	<BEGIN> ***. Show the diagram and the text explanation.	<BEGIN> How should I handle not being able to keep track of the characters while reading "War and Peace"? Show the diagram and the text explanation.
Multi-view News Generation	Please output interleaved images and texts for required reports: <BEGIN> ***	Please output interleaved images and texts for required reports: <BEGIN> How can the announcement by the United States of additional military aid to Ukraine be reported from multiple perspectives?
Dynamic Sports Event Analysis	<BEGIN> ***. Please recreate the scenes with text and images.	<BEGIN> In the third round of La Liga 2024, Real Madrid drew 1-1 away against Las Palmas. Please recreate the moment of the goals with text and images.
Interactive Historical Interpretation	<BEGIN> ***. Please provide a brief history of this event using images.	<BEGIN> Are you aware of the Pearl Harbor incident? Please provide a brief history of this event using images.
Unsolved Mysteries Exploration	Please answer the question with image and text explanation: <BEGIN> ***	Please answer the question with image and text explanation: <BEGIN> Could you help deduce how the Mycenaean civilization was destroyed?
Dream Analysis and Scene Reconstruction	I had a dream. Please help me visualize my dream into an image, and analyze why I had this dream, what are the implications and meanings? This is the content of my dream: <BEGIN> ***	<BEGIN> I had a dream. Please help me visualize my dream into an image, and analyze in words why I had this dream, including any implications and meanings. Here is the content of my dream: I dreamt of meeting a girl I know at the place where we first met
Multimodal Biological Reasoning	<BEGIN> ***. Are there any more photos of this species? <image>	<BEGIN> May I ask what species of fish this is? Are there any more photos of this species? <image>
Multimodal Landscape Reasoning	<BEGIN> ***. Could you provide me with more photos of this and introduce them to me? <image>	<BEGIN> Which city are these photos from? Could you provide me with more landscape photos of this city and introduce them to me? <image>
Multimodal Analogy Reasoning	Please answer this question with image and text explanation: <BEGIN> ***	Please answer this question with image and text explanation: <BEGIN> What should be filled in the question mark to make it exhibit a certain regularity? <image>
Interactive Jigsaw Puzzle	<BEGIN> ***. Show the resulting image with the corresponding text explanation.	<BEGIN> Here are some puzzle pieces. Please assemble them into a complete picture. Show the resulting image with the corresponding text explanation. <image>
Interactive Novel View Synthesis	<BEGIN> *** Please draw the picture and give descriptions.	<BEGIN> This is a pear slice. Its appearance features are:. Can you guess what it looks like from the side? Please draw the picture and give descriptions. <image>

Task	Template	Example
Interactive Multi-concept Image Composition	<BEGIN> ***. Please summarize all the content in one image and write a blog post. <image>	<BEGIN> This is a collection of four Christmas smoothies. Please summarize all the content in one image and write a blog post. <image>
Interactive Film and Television Recommendation	Please output recommendations in the form of the poster and the corresponding introduction: <BEGIN> ***	Please output recommendations in the form of the poster and the corresponding introduction: <BEGIN> Could you recommend some Indian dramas to me?
Interactive Goods Recommendation	Please output recommendations in the form of images and give the corresponding introduction: <BEGIN> ***	Please output recommendations in the form of images and give the corresponding introduction: <BEGIN> Are there any throw pillows you can recommend?
Interactive Food Recommendation	<BEGIN> ***. Please provide the information with images and text.	<BEGIN> What are some recommended dishes in Yibin, Sichuan? Please provide the information with images and text.
Business Scenarios Brainstorming	<BEGIN> I want to start a business. Please brainstorm with me about some ways to start a business and help me figure it out. *** Please output brainstorming results with images and explanations.	<BEGIN> I want to start a business. Please brainstorm with me about some ways to start a business and help me figure it out. Please analyze the long-term development trends of automotive braking technology for me. Please output brainstorming results with images and explanations.
Academic Scenarios Brainstorming	<BEGIN> What/Why/How ***? Please also show an illustration.	<BEGIN> What are the steps involved in the synthesis of Metal-Organic Frameworks (MOFs) using the hydrothermal method? Please also show an illustration.
Multimodal Action Anticipation	In this task, you are given the first part of an activity with both text and an image, and you need to complete the subsequent action parts of the activity by generating text and images that are natural continuation of the given first part. The input interleaved content is: <BEGIN> ***	In this task, you are given the first part of an event with both text and an image, and you need to complete the subsequent parts of the event by generating text and images that are natural continuation of the given first part. The input interleaved content is: <BEGIN> A boy is trying to go through the security gate at the airport. <image>
Visual Traffic Forecasting	<BEGIN> What will the traffic conditions ***? Please provide an explanation and present it in the form of images.	<BEGIN> What will the traffic conditions be like near Fuxing Road in Shenzhen in an hour? Please provide an explanation and present it in the form of images. <image>
Interactive Remote Sensing Image Rendering	<BEGIN> *** Also give interleaved text explanations for generated images.	<BEGIN> Please generate a remote sensing satellite image of the area based on my geographical photo. Also give interleaved text explanations for generated images. <image>
Interactive Street View Image Rendering	<BEGIN> *** Also give interleaved text explanations for generated images.	<BEGIN> Please generate a panorama of the area based on my remote sensing satellite image. Also give interleaved text explanations for generated images. <image>
Urban Planning and Development Simulation	Please output the scheme in the form of both the image and the text explanation to meet the requirements: <BEGIN> ***	Please output the scheme in the form of both the image and the text explanation to meet the requirements: <BEGIN> In accordance with this planning diagram, please design a final rendering. <image>
Plog and Social Media Content Generation	<BEGIN> ***. Could you create a social media post with text and images?	<BEGIN> After finishing Jia Pingwa's "Comfortably Alone," I am deeply moved and want to post something on social media but don't know how to phrase it. Could you help me create a post with text and images for my use?

Chat with Memes	You are a funny chatbot that responds to my small talk. Please output meme images to interact with me and chat with me. <BEGIN> ***	You are a funny chatbot that responds to my small talk. Please output meme images to interact with me and chat with me. <BEGIN> We'll be traveling in three weeks! <image>
Interactive Virtual Try-on	<BEGIN> Please generate a visualization of ***, and provide an evaluation of the try-on effect.	<BEGIN> Please generate a visualization of how the clothing looks when worn, based on the photos of the clothing and the model I provided, and provide an evaluation of the fitting effect. <image>
Multimodal Dressing Suggestion	<BEGIN> ***. Please provide the information in both text and images.	<BEGIN> What are some outfit suggestions for women traveling in the summer? Please provide the information in both text and images.
Fashion Trend Forecasting	<BEGIN> What are the *** trend in the upcoming ***? Please provide the information in both text and images.	<BEGIN> What are the design elements for men's shoes in the upcoming autumn and winter? Please provide the information in both text and images.
Multimodal Recipe Generation	<BEGIN> How to prepare ***? Please provide the steps in a detailed format with images and text.	<BEGIN> How to prepare this type of soy sauce boiled pomfret: Please provide the steps in a detailed format with images and text. <image>
Multimodal Cooking Assistant	Please output instructions in interleaved images and texts: <BEGIN>	Please output instructions in interleaved images and texts: <BEGIN> The batter I made for the egg burger doesn't taste good. How can I make it taste better? <image>
Interactive Tutorial Generation	Please show me the steps of the tutorial with interleaved images and text: <BEGIN>	Please show me the steps of the tutorial with interleaved images and text: <BEGIN> Please tell me how to seal or protect the finish of painted wood.
Interactive Science Popularization	<BEGIN> What is ***? Please explain with illustrations and text.	<BEGIN> What is the Doppler Effect? Please explain with illustrations and text.
Fitness and Health Consulting	Please give answers in interleaved images and texts: <BEGIN> ***	Please give answers in interleaved images and texts: <BEGIN> Please tell me what issues I need to pay attention to when engaging in walking exercise.
Autonomous Driving and In-door Navigation	<BEGIN> You're an embodied AI that captures your surroundings through a camera. These are images captured in the past and present. What will the proceeding image possibly be in the next frame?	<BEGIN> Assume that you are an embodied-AI agent and perceive the surroundings through a camera. You were presented with a series of three images from the past to present. Try to determine what the proceeding image could possibly be. <image>

You are an impartial judge in multimodal content evaluation. Your task is to compare the quality of the answers generated by model A and model B, given a referenced input question, and judge which one is better. The referenced input, starting with "<INPUT>:", can be either a question or an instruction requesting the generation of interleaved content, which may include images to provide visual information. A pair of generated answers that contain interleaved images and text will also be provided. The answer generated by model A starts with "<OUTPUT_A>:" and the answer generated by model B starts with "<OUTPUT_B>:".

Evaluation Criteria:

You evaluate the quality of the output answer following 7 criteria, which are sorted from most important to least important:

1. **Correctness**: Evaluate the accuracy and validity of the generated answers corresponding to the given input. This includes if the text exists and if the image exists. Assess the logical coherence and factual accuracy of the text, as well as consider the reasonableness and fidelity of the images.

2. **Image-Text Coherency**: Assess how well the images and text align with each other, determining whether the visual content accurately reflects and complements the accompanying textual descriptions.

3. **Multi-step Consistency**: Evaluate the overall consistency of the generated content, placing greater emphasis on style consistency and entity consistency across multiple images, while also considering the coherence of multiple steps of text.

4. **Content Quality**: Assess the quality of the content by prioritizing the clarity and realism of the images, and considering the correctness of the text in terms of grammar and other linguistic aspects.

5. **Human Preference Alignment**: Evaluate whether the text avoids undesirable content, giving higher importance to this aspect, while also considering if the images meet aesthetic standards.

6. **Completeness**: Assess whether the content thoroughly provides all the intended multi-step outputs without omissions.

7. **Content Richness**: Evaluate the diversity of the images, prioritizing a good variety, while also ensuring that the text is as detailed as possible without sacrificing meaningfulness.

Instructions:

- As you can accept image inputs directly, thoroughly analyze both the images and text during evaluation. Ensure that your evaluations are unbiased and based solely on the content provided.
- When determining the final verdict, prioritize the criteria listed earlier.
- After evaluating all criteria, compare the total performance and provide a final verdict on which model performed better overall.
- Please directly output "A is better" or "B is better" or "Tie (A is better)" or "Tie (B is better)".

Data to Review:

Figure 12. The system prompt for using GPT-4o as a judge to compare outputs from two interleaved generation methods.

Refined Prompt

You are an impartial judge in multimodal content evaluation. Given a referenced input question (marked by <INPUT>:), your task is to compare the quality of the answers generated by model A (marked by <OUTPUT_A>:) and model B (marked by <OUTPUT_B>:), and judge which one is better.

Data to Review:

Please directly output "A is better" or "B is better" or "Tie(A is better)" or "Tie(B is better)":

Unrefined Prompt

You are an impartial judge in multimodal content evaluation. Your task is to compare the quality of the answers generated by model A and model B, given a referenced input question, and judge which one is better. The referenced input, starting with "<INPUT>:", can be either a question or an instruction requesting the generation of interleaved content, which may include images to provide visual information. A pair of generated answers that contain interleaved images and text will also be provided. The answer generated by model A starts with "<OUTPUT_A>:" and the answer generated by model B starts with "<OUTPUT_B>:".

Evaluation Criteria:

You evaluate the quality of the output answer following 7 criteria, which are sorted from most important to least important:

1. **Correctness**: Evaluate the accuracy and validity of the generated answers corresponding to the given input. This includes if the text exists and if the image exists. Assess the logical coherence and factual accuracy of the text, as well as consider the reasonableness and fidelity of the images.

2. **Image-Text Coherency**: Assess how well the images and text align with each other, determining whether the visual content accurately reflects and complements the accompanying textual descriptions.

3. **Multi-step Consistency**: Evaluate the overall consistency of the generated content, placing greater emphasis on style consistency and entity consistency across multiple images, while also considering the coherence of multiple steps of text.

4. **Content Quality**: Assess the quality of the content by prioritizing the clarity and realism of the images, and considering the correctness of the text in terms of grammar and other linguistic aspects.

5. **Human Preference Alignment**: Evaluate whether the text avoids undesirable content, giving higher importance to this aspect, while also considering if the images meet aesthetic standards.

6. **Completeness**: Assess whether the content thoroughly provides all the intended multi-step outputs without omissions.

7. **Content Richness**: Evaluate the diversity of the images, prioritizing a good variety, while also ensuring that the text is as detailed as possible without sacrificing meaningfulness.

Instructions:

- As you can accept image inputs directly, thoroughly analyze both the images and text during evaluation. Ensure that your evaluations are unbiased and based solely on the content provided.
- When determining the final verdict, prioritize the criteria listed earlier.
- After evaluating all criteria, compare the total performance and provide a final verdict on which model performed better overall.
- Please directly output "A is better" or "B is better" or "Tie (A is better)" or "Tie (B is better)".

Data to Review:

Figure 13. The system prompts for using MLLMs as a judge to compare outputs of two interleaved generation methods.

You are an impartial judge in multimodal content evaluation. Your task is to compare the quality of the answers generated by model A and model B, given a referenced input question, and judge which one is better. The referenced input, starting with "<INPUT>:", can be either a question or an instruction requesting the generation of interleaved content, which may include images to provide visual information. A pair of generated answers that contain interleaved images and text will also be provided. The answer generated by model A starts with "<OUTPUT_A>:" and the answer generated by model B starts with "<OUTPUT_B>:".

Image Representation:

- The image for the step *i* will be enclosed by the symbols `<IMG_i>` and `</IMG_i>`.
- The images are numbered sequentially from **0 to N** (including the images in input, the images generated by model A and the images generated by model B).

Evaluation Criteria:

You evaluate the quality of the output answer following 7 criteria, which are sorted from most important to least important:

1. **Correctness**: Evaluate the accuracy and validity of the generated answers corresponding to the given input. This includes if the text exists and if the image exists. Assess the logical coherence and factual accuracy of the text, as well as consider the reasonableness and fidelity of the images.

2. **Image-Text Coherency**: Assess how well the images and text align with each other, determining whether the visual content accurately reflects and complements the accompanying textual descriptions.

3. **Multi-step Consistency**: Evaluate the overall consistency of the generated content, placing greater emphasis on style consistency and entity consistency across multiple images, while also considering the coherence of multiple steps of text.

4. **Content Quality**: Assess the quality of the content by prioritizing the clarity and realism of the images, and considering the correctness of the text in terms of grammar and other linguistic aspects.

5. **Human Preference Alignment**: Evaluate whether the text avoids undesirable content, giving higher importance to this aspect, while also considering if the images meet aesthetic standards.

6. **Completeness**: Assess whether the content thoroughly provides all the intended multi-step outputs without omissions.

7. **Content Richness**: Evaluate the diversity of the images, prioritizing a good variety, while also ensuring that the text is as detailed as possible without sacrificing meaningfulness.

Instructions:

- As you can accept image inputs directly, thoroughly analyze both the images and text during evaluation. Ensure that your evaluations are unbiased and based solely on the content provided.
- When determining the final verdict, prioritize the criteria listed earlier.
- After evaluating all criteria, compare the total performance and provide a final verdict on which model performed better overall.
- Please directly output "A is better" or "B is better" or "Tie(A better)" or "Tie(B better)".

Data to Review:

Figure 14. The system prompt for obtaining detailed scores from GPT-based evaluators. Brief explanations are also required to support further performance analysis of different models.

The End-to-end Interleaved Generator models, on the other hand, represent a significant shift towards multimodal generation of interleaved image-text content. MiniGPT-5, GILL, NExT-GPT and SEED-LLaMA are designed to generate interleaved text and images in a single unified process, eliminating the need for intermediate stages. This integrated approach not only reduces latency but also improves the alignment and contextual relevance between text and images. It is noted that Anole is the only model that can directly output multi-step image-text content, whereas it is fine-tuned based on the powerful capabilities of Chameleon [30]. However, the open-sourced Chameleon only releases the weights for text generation, and the weights for image generation are withheld by randomizing the corresponding parameters.

Collectively, these models demonstrate diverse strategies for addressing challenges in interleaved multimodal generation, from modular pipelines to unified architectures. This taxonomy allows comprehensive evaluation and analysis of potential directions for developing MLLMs. We detail each model by category below:

- **Integrated Pipelines:**
  - **GPT-4o+DALL-E·3** [3, 24]: This pipeline leverages GPT-4o [24] to generate text and captions for the desired image generation. The captions are subsequently fed into DALL-E·3[3] to produce the corresponding images. The final output combines the text and images in their original order, enabling multimodal content generation through a staged process.
  - **Gemini1.5+Flux** [4, 31]: This method integrates Gemini1.5 Pro for text generation with Flux-schnell for fast and efficient image generation. The pipeline emphasizes high-quality and coherent text-to-image alignment through a structured two-step process.
- **Two-stage Generators:**
  - **Emu2** [29]: Emu2 is a 37B MLLM with multimodal generation capabilities. The pretrained Emu2 is fine-tuned separately on conversational and image data, enabling it to function as Emu2-Chat for multimodal understanding and Emu2-Gen for image generation. We implement Emu2-Chat and Emu2-Gen in a two-stage pipeline to ensure seamless interleaved outputs.
  - **SEED-X** [15]: SEED-X is a unified multimodal foundation model that integrates multi-granularity visual comprehension and generation. We also implement this model in a two-stage pipeline approach, generating interleaved text and images in separate stages, since the prompts for instructing the model to comprehend multimodal input and generate image tokens are different.
  - **Show-o** [39]: Show-o is a unified Transformer model combining autoregressive and diffusion approaches to flexibly handle multimodal understanding and generation tasks. We implement Show-o, which adopts a similar two-stage generation approach but separately

producing interleaved multimodal content step by step.
  - **Emu3** [34]: Emu3 is one of the latest MLLMs trained in next-token prediction, capable of generating high-quality images, text, and videos by tokenizing multimodal sequences into a discrete space and training a single transformer, achieving superior performance over various SOTA models such as SDXL and LLaVA-1.6. We implement Emu3-Chat (finetuned on multimodal understanding data) and Emu3-Gen (finetuned on visual generation data) in a two-stage pipeline to ensure seamless interleaved outputs.
  - **VILA-U** [38]: VILA-U is a unified foundation model integrating video, image, and language understanding and generation through a single autoregressive next-token prediction framework, achieving near SOTA performance in various multimodal tasks. We implement VILA-U in a similar two-stage generation approach since it has separate multimodal understanding and image generation abilities.
- **End-to-End Generators:**
  - **MiniGPT-5** [43]: MiniGPT-5 directly generates interleaved text and images in an end-to-end manner. It combines MiniGPT-4 and Stable Diffusion, using "generative vokens" to seamlessly connect the textual and visual domains for efficient and coherent generation. Its seamless integration enables efficient and coherent multimodal generation without intermediate steps. In particular, MiniGPT-5 has two different versions trained on VIST and MMDialog, respectively. We name the version trained on VIST as MiniGPT-5 because it is the most widely used. The version trained on MMDialog is named MiniGPT-5MMD.
  - **GILL** [19]: GILL fuses frozen text-only LLMs with pretrained visual models using a mapping network. It maps text hidden states from the pretrained LLM to map text hidden states into the embedding space of an image generation model, allowing multimodal generation.
  - **NExT-GPT** [36]: NExT-GPT is an end-to-end MLLM capable of processing and generating text, images, videos, and audio in any combination. We implement it by removing the video and audio generation flow and the remaining text and image generation abilities. It can directly output interleaved multimodal content through its streamlined architecture.
  - **SEED-LLaMA** [14]: SEED-LLaMA integrates text and image generation into a unified framework through the SEED tokenizer, enabling both comprehension and generation of text and images. It offers a direct end-to-end solution for creating interleaved multimodal content.
  - **Anole** [10]: Anole is an end-to-end interleaved generation model fine-tuned on Chameleon, leveraging pretrained weights of Chameleon to produce high-quality interleaved text, complemented with coherent images

generated by optimizing image token logits in the output layer. It is the only available model that can directly output multistep image-text content.

D.2. Implementation Details

The experiments were conducted using a total of 24 A100 80G GPUs, with 8 GPUs dedicated to training IntJudge. We explored different large multimodal language models (MLLMs), including InternLM-XComposer2.5 (InternLMX2.5) and Qwen2-VL, ultimately selecting Qwen2-VL-7B as the foundational model for training IntJudge to achieve an optimal balance between efficiency and accuracy. The training process involved LoRA-based parameter-efficient fine-tuning based on LLaMA-Factory. To optimize training performance, DeepSpeed and FlashAttention-2 are adopted. We define a cutoff length of 16,240 tokens for inputs. We use a per-device batch size of 1, gradient accumulation steps of 8, a learning rate of 1.0e-4, a cosine learning rate schedule, and 20 epochs with BF16 mixed-precision enabled. The evaluation process involved sampling comparison pairs. Specifically, we conducted sampling rounds to obtain a total of E distinct battle pairs for each data instance. The sampling round value E was set to 2, resulting in 4,320 battle pairs being formed for comparison.

D.3. Experimental Results on New Models

We present more main experimental results in Table 12. The new models Emu3, VILA-U and MiniGPT-5MMD are also evaluated by Human, GPT-based, and IntJudge-based evaluators and compared with 10 established baseline models using the win rate metrics. The results of methods are ranked by their performance on FDT metric evaluated by Human. Table 12 shows a clear hierarchy in model performance, with Human and GPT-4o+DALL-E3 leading across all metrics.

A closer look at the results reveals a consistency in rankings across evaluators. For instance, GPT-4o+DALL-E3 consistently secures second place in Human Evaluation and IntJudge Evaluation. Conventional end-to-end models, such as MiniGPT-5 and GILL, struggle to match the quality of their competitors, highlighting their limitations in generating contextually relevant and diverse outputs. However, GPT Evaluation shows a clear preference for outputs by GPT-4o+DALL-E3. It is verified that GPT-based judgments are not objective enough due to the inherent bias. In contrast, our proposed IntJudge shows better alignment with human judgments, supporting the reliability of IntJudge as an effective evaluation framework.

Different evaluation metrics also offer more details about model performance. The FDT metric, which forces a decision in tie cases, highlights the dominance of Human and GPT-4o+DALL-E3. However, metrics that account for ties more flexibly, such as "w/ Tie (0)" and "w/ Tie (.5)," elevate end-to-end models like VILA-U and Emu3, suggesting

that these models produce outputs that, while not always definitive winners, are frequently competitive. This distinction underscores the importance of using diverse metrics to capture various dimensions of model performance.

The new two-stage models show promising results, with VILA-U standing out for its balanced performance across all metrics, making it a reasonable option for general interleaved image-text tasks. MiniGPT-5MMD (finetuned on MMDialog) shows slight improvements over its variant MiniGPT-5 (finetuned on VIST), indicating progress but still trailing behind the latest models. Meanwhile, Emu3 performs well under specific metrics, such as "w/ Tie (.5)," showing the potential to generate tie-worthy outputs with a certain quality.

The results also highlight the challenges faced by conventional end-to-end models, such as NExT-GPT, and GILL, which consistently underperform. These models reveal the inherent difficulty in achieving coherence and contextual relevance in interleaved generation tasks. Though Anole achieved a decent ranking as a representative end-to-end model, more advanced end-to-end models are needed for better visual generation quality.

Overall, the experimental results validate the effectiveness of IntJudge as a reliable evaluator, demonstrating its consistency with human judgments. The analysis underscores the strengths of integrated generation pipelines such as GPT-4o+DALL-E3 and Gemini1.5+Flux. and identifies opportunities for improvement in two-stage and end-to-end models. Looking forward, expanding the training dataset, enhancing model architectures and improving the evaluation methods will all be critical in driving further progress in open-ended interleaved image-text generation, pushing the boundaries of multimodal learning research.

D.4. Main Results Breakdown

Fig. 15 presents the win rates of 14 interleaved generation methods across 23 meta-topics, evaluated solely through human evaluations. The methods are evaluated using four distinct metrics: Force Dividing Tie (FDT), Without Tie, With Tie (0), and With Tie (0.5). The results are presented using histogram figures, which provide a clear visual comparison of model performance across different topic scenarios. For example, SOTA models like GPT-4o+DALL-E3, Emu3, and VILA-U consistently ranked high in categories like "Storybook Creation," "Graph Generation," and "2D Image Reasoning," showcasing their superior capabilities in generating coherent interleaved content. Conversely, models like MiniGPT-5, NExT-GPT, and GILL struggled across most tasks, especially in areas such as "Healthcare Tasks," "Multimodal Time Series Forecasting," and "Educational Tasks," indicating a need for improved contextual understanding and generation capabilities. Training on larger datasets that include more domain knowledge may mitigate these issues and improve their interleaved generation performance.
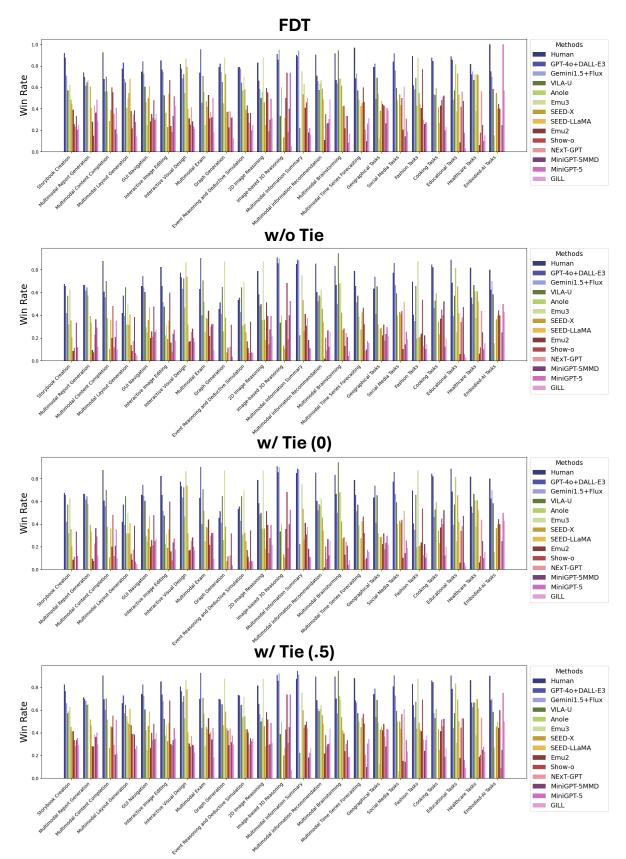
Figure 15. The win rates of 14 interleaved generation methods across 23 meta-topics.

Method	Human Evaluation				GPT Evaluation				IntJudge Evaluation			
	FDT	w/o Tie	w/ Tie (0)	w/ Tie (.5)	FDT	w/o Tie	w/ Tie (0)	w/ Tie (.5)	FDT	w/o Tie	w/ Tie (0)	w/ Tie (.5)
Human	83.94%	86.50%	70.78%	79.87%	82.76%	83.09%	82.27%	82.76%	85.65%	89.11%	72.62%	81.87%
GPT-4o+DALL-E3	78.20%	80.73%	66.17%	75.19%	86.33%	86.60%	86.23%	86.44%	83.24%	86.20%	71.46%	80.01%
Gemini1.5+Flux	66.67%	66.95%	51.97%	63.16%	73.39%	73.38%	72.75%	73.18%	66.11%	67.92%	49.58%	63.08%
VILA-U	62.10%	62.34%	61.57%	62.19%	49.47%	49.55%	49.29%	49.56%	68.66%	58.58%	36.94%	55.41%
Anole	52.72%	53.10%	38.96%	52.28%	53.25%	53.06%	52.60%	53.04%	56.33%	52.77%	33.85%	51.78%
Emu3	54.05%	55.24%	52.25%	54.95%	47.19%	47.27%	46.74%	47.30%	54.01%	54.48%	39.04%	53.21%
SEED-X	53.25%	52.03%	38.55%	51.51%	56.46%	56.63%	55.63%	56.51%	53.76%	54.32%	36.15%	52.88%
SEED-LLaMA	44.43%	42.47%	30.76%	44.54%	42.33%	42.13%	41.68%	42.22%	46.43%	45.49%	28.21%	47.20%
Emu2	40.31%	36.64%	24.78%	40.97%	42.49%	42.43%	41.52%	42.60%	36.60%	31.84%	19.36%	38.96%
NExT-GPT	33.59%	27.74%	18.76%	34.95%	24.81%	24.62%	24.27%	24.97%	34.08%	25.94%	15.39%	35.72%
Show-o	37.47%	35.97%	24.57%	40.42%	33.21%	32.81%	32.26%	33.10%	33.65%	24.22%	13.59%	35.53%
MiniGPT-5MMD	32.26%	32.04%	30.74%	32.77%	32.59%	32.47%	32.25%	32.59%	28.98%	25.30%	14.84%	35.51%
MiniGPT-5	31.47%	28.59%	19.72%	35.24%	31.18%	30.98%	30.66%	31.18%	24.65%	15.65%	9.08%	30.07%
GILL	25.96%	20.82%	14.33%	29.91%	31.47%	31.25%	30.70%	31.58%	23.23%	16.80%	10.35%	29.54%

Table 12. Comparison of model win rates evaluated by human, GPT-4o, and our IntJudge under FDT and different tie metrics. FDT: Force Dividing Tie metric. w/o Tie: Non-tie case. w/ Tie (0) and w/ Tie (.5): Count a tie as 0 and 0.5 wins for a model in a battle, respectively.

D.5. More Pairwise Model Performance

Figure 16 presents more heatmaps that illustrate the pairwise model performance evaluated by different evaluators, including Human, GPT, and IntJudge. These heatmaps provide a visual representation of the comparative strengths and weaknesses of each model across multiple metrics, such as Force Dividing Tie (FDT) and different approaches to handling tie cases (without ties, ties as zero, and ties as 0.5). By examining these heatmaps, we gain a clearer understanding of how well each model fares against others, diving deeper into performance consistency and discrepancies across evaluators.

D.6. More Ablations on Sampling Size

Figure 17 illustrates the results of additional ablation studies focusing on the effect of sampling size on model performance. The figure compares win rates across different evaluators, including Human, GPT, and IntJudge, under various metrics such as Force Dividing Tie (FDT) and different methods for treating ties (without ties, ties as zero, and ties as 0.5). These ablation studies are crucial for understanding the impact of sampling on the robustness of model comparisons and provide insights into how sampling variations influence the ranking consistency among different evaluators. Most importantly, the results help validate the stability of our evaluation framework.

D.7. Case Study

Here, we present case studies of interleaved content generation, which involve 14 distinct models competing in the OpenING tasks. From these models, 23 model battle pairs are formed, and each pair engages in a battle over a representative task from the 23 meta-topics in our OpenING benchmark. The results of these battles are shown in Fig. 19. A human judge awards a gold medal to the model that wins

absolutely over its competitor in a pair. In cases where the generated content by two competing models is of similar qualities, the judge awards use tie metrics to determine and award the silver medal to the slightly better-performing model. The human judgments are based on the following criteria, which include the eight typical errors these models tend to commit in the interleaved content generation:

1. **No-Text or No-Image**: A model fails to generate text or images or both when explicitly instructed or expected to do so.
2. **Factual Error**: A model may present incorrect or made-up information, such as referencing nonexistent studies, authors, or events. Additionally, it might fail to follow instructions, examples include refusing to provide answers, misinterpreting the input and generating irrelevant responses, displaying reasoning mistakes, or producing inaccurate images.
3. **Content Incoherent**: The generated content lacks coherence with the input or is inconsistent across multiple outputs, typically in style or entities
4. **Offensive Content**: Offensive content is defined as materials that include distorted or disturbing imagery or scenes that likely cause discomfort or distress to readers. It also includes content that raises safety concerns or violates safety guidelines, such as depictions of violence, harm, hate speech, adult content, illegal activities, dangerous behaviors, and unethical actions. A model may also fail to consider cultural nuances or sensitivities.
5. **Image-Text Inconsistent**: This error occurs when the images do not semantically align with the corresponding text, leading to confusion or misinterpretation.
6. **Poor Image Quality**: The generated images have low quality, such as being completely blank or blacked-out,
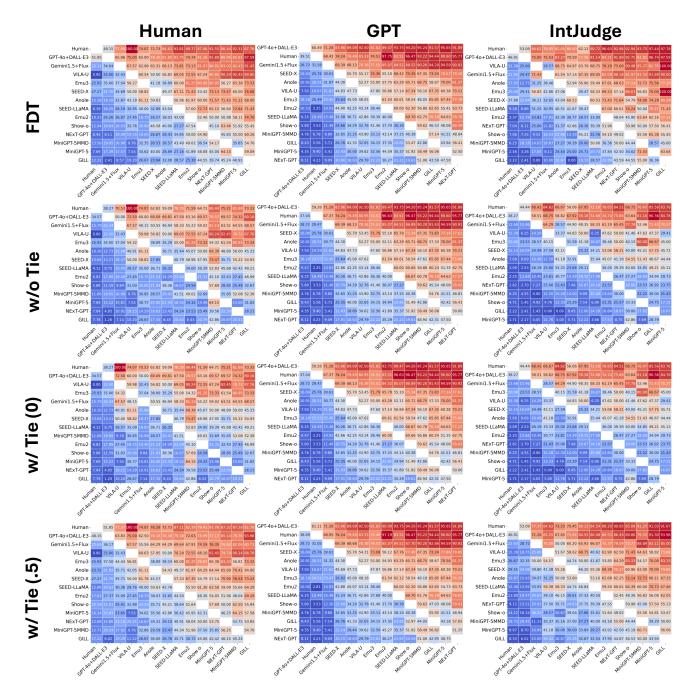
Figure 16. Win rate matrices of 14 interleaved genration methods, evaluated by Human, GPT-4o, and our IntJudge, respectively.

blurry, poorly laid out, and lacking realism or aesthetic appeal.

7. **Poor Text Quality**: The generated texts are of low quality, including nonsensical mumbling, grammatical errors that impair readability and understanding, and being too short to convey meaningful information.

8. **Incomplete Response**: A model abruptly stops generating a textual response, or a model outputs an incomplete or truncated response.

E. Finetuning MLLMs on OpenING

We present the extended experimental results of training MiniGPT-5 on the Dev Set of OpenING and testing the fine-tuned model on the Test Set of OpenING. The objective is to verify if finetuning on the specific data of OpenING can im-
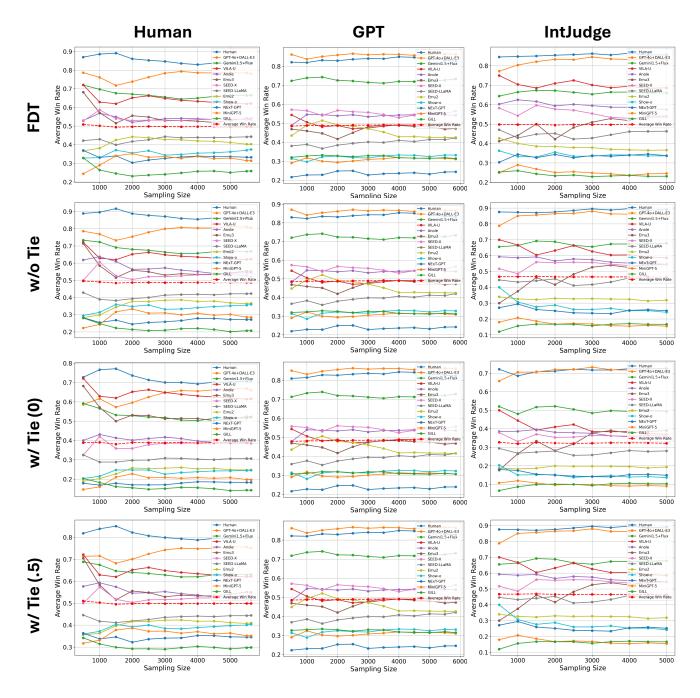
Figure 17. Win rate curves with respect to different sampling sizes.

prove the performance of interleaved generation tasks. The Dev Set of OpenING can offer a set of 3,000 training samples that align with the diverse unique tasks. The MiniGPT-5 model was finetuned using the Dev Set for 5 epochs with a learning rate of 2e-5, utilizing an Adam optimizer. To enhance training stability, Adam epsilon of 1e-8 was applied. The model training incorporated mixed-precision computations to speed up the training process. The results are evaluated on the Test Set of OpenING using IntJudge.

In Table 13, the performance of MiniGPT-5OpenING, the finetuned version of MiniGPT-5, is compared against other state-of-the-art models and the original MiniGPT-5 baselines (MiniGPT-5 is finetuned on VIST and MiniGPT-5MMD is finetuned on MMDialog). We set E to 1 and randomly sampled 2,160 samples for this efficient evaluation. The evaluation metrics include four scenarios: Force Dividing

Model	FDT	w/o Tie	w/ Tie (0)	w/ Tie (.5)
Human	84.66%	86.01%	75.46%	81.60%
Gemini1.5+Flux	73.44%	73.15%	61.72%	69.53%
VILA-U	62.50%	60.14%	41.50%	57.00%
MiniGPT-5OpenING	60.24%	63.76%	44.71%	59.65%
Emu3	56.02%	55.20%	36.13%	53.40%
SEED-X	54.23%	54.67%	41.80%	53.57%
Emu2	47.10%	39.33%	25.36%	43.12%
Anole	41.56%	43.10%	32.47%	44.81%
SEED-LLaMA	41.33%	38.14%	24.67%	42.33%
GILL	36.25%	32.04%	20.62%	38.44%
Show-o	35.76%	30.53%	19.21%	37.75%
MiniGPT-5	31.54%	26.37%	16.11%	35.57%
MiniGPT-5MMD	28.19%	23.44%	14.71%	33.33%

Table 13. Model Win Rates evaluated by IntJudge. Representative models for each type of methods are chosen to be compared with MiniGPT-5OpenING, which is a MiniGPT-5 version finetuned on the Dev Set of OpenING. FDT: Force Dividing Tie metric. w/o Tie: Non-tie case. w/ Tie (0) and w/ Tie (.5): Count a tie as 0 and 0.5 wins for a model in a battle, respectively.

Tie (FDT), Without Tie (w/o Tie), With Tie counted as 0 (w/ Tie (0)), and With Tie counted as 0.5 (w/ Tie (.5)).

The results highlight that MiniGPT-5OpenING achieves significant improvements over the baseline MiniGPT-5 models across all metrics. For example, in the Without Tie (w/o Tie) scenario, the finetuned model shows a substantial 37.39% relative improvement over the MiniGPT-5 baseline. These findings confirm that training on a specialized interleaved image-text dataset such as the Dev Set of OpenING enhances the model with better contextual understanding and alignment capabilities for generating coherent interleaved image-text content. Further studies are ongoing to improve the performance of SOTA models.

F. Limitations of This Study

Although OpenING advances interleaved image-text generation evaluation, several limitations remain for improvement. First, despite covering diverse tasks (56 tasks across 23 meta-topics), some real-world scenarios remain underrepresented or oversimplified, potentially limiting the generalizability to practical applications. Tasks requiring fine-grained understanding or multi-step reasoning need to be supplied to capture real-world needs. Second, although the IntJudge model improves alignment with human evaluations, its generalizability remains limited by the diversity and quality of training data. The benchmark's reliance on human-annotated data to establish ground truth and train judge models is labor-intensive and expensive. While our Reference-Augmented Generation (RAG) approach helps scale training data, manual annotations remain a critical component for ensuring quality and alignment with human preferences. Furthermore, the computational resources required for training and deploying IntJudge present scalability challenges, potentially

limiting accessibility for researchers with fewer resources.

In addition, current interleaved image-text generation methods still struggle with producing high-quality, coherent interleaved content, particularly in multi-step tasks that require maintaining consistency across generated images and text. Issues like content incoherence, poor image quality, and mismatches between generated text and images persist across evaluated models, particularly in end-to-end approaches. To tackle these issues, more advanced MLLMs trained with a large-scale interleaved image-text dataset are to be investigated. Moreover, while our IntJudge demonstrates significant advantages over GPT-based evaluators, we recognize two aspects for improvement: 1) Potential biases may rise from crowdsourced data (e.g., human preferences in aesthetic judgment), and 2) The current foundation model of IntJudge mainly supports English and Chinese. Building a sufficiently comprehensive, diverse, and representative dataset is expected to greatly promote the development of multimodal generation.

These limitations underscore the need for continued development of more diverse datasets and more robust evaluation frameworks to address the complexities of interleaved generation evaluation, enabling more practical interleaved image-text generation methods and pushing forward the boundary of future MLLMs. Future work may benefit from diversifying data sources and cross-cultural annotations, expanding multilingual capabilities, and implementing debiasing techniques to improve fairness.

References

[1] Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. BoQ: A place is worth a bag of learnable queries. In *Proceed-*

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 17794–17803, 2024. 6

[2] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. *CVPR*, 2022. 6

[3] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science*, 2(3):8, 2023. 9, 19

[4] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. Accessed: 2024-11-05. 9, 19

[5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 5

[6] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 7

[7] Yitao Cai, Huiyu Cai, and Xiaojun Wan. Multi-modal sarcasm detection in Twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019. 7

[8] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. MLLM-as-a-Judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Proceedings of the International Conference on Machine Learning*, 2024. 2

[9] Dongping Chen, Yue Huang, Siyuan Wu, Jingyu Tang, Liuyi Chen, Yilin Bai, Zhigang He, Chenlong Wang, Huichi Zhou, Yiqiang Li, Tianshuo Zhou, Yue Yu, Chujie Gao, Qihui Zhang, Yi Gui, Zhen Li, Yao Wan, Pan Zhou, Jianfeng Gao, and Lichao Sun. Gui-world: A dataset for gui-oriented multimodal llm-based agents, 2024. 5

[10] Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. *arXiv preprint arXiv:2407.06135*, 2024. 9, 19

[11] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018. 7

[12] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017. 7

[13] Zhengcong Fei, Zekang Li, Jinchao Zhang, Yang Feng, and Jie Zhou. Towards expressive communication with internet memes: A new multimodal conversation dataset and benchmark. *arXiv preprint arXiv:2109.01839*, 2021. 7

[14] Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making llama see and draw with seed tokenizer. *arXiv preprint arXiv:2310.01218*, 2023. 9, 19

[15] Yuying Ge, Sijie Zhao, Chen Li, Yixiao Ge, and Ying Shan. Seed-data-edit technical report: A hybrid dataset for instructional image editing. *arXiv preprint arXiv:2405.04007*, 2024. 5, 9, 19

[16] Yulia Gryaditskaya, Mark Sypesteyn, Jan Willem Hoftijzer, Sylvia Pont, Frédo Durand, and Adrien Bousseau. Opensketch: A richly-annotated dataset of product design sketches. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 38, 2019. 5

[17] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6047–6056, 2018. 7

[18] Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Pnp inversion: Boosting diffusion-based editing with 3 lines of code. *International Conference on Learning Representations (ICLR)*, 2024. 5

[19] Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. Generating images with multimodal language models. In *Proceedings of the Advances in Neural Information Processing Systems*, 2024. 9, 19

[20] Zhikai Li, Xuewen Liu, Dongrong Fu, Jianquan Li, Qingyi Gu, Kurt Keutzer, and Zhen Dong. K-sort arena: Efficient and reliable benchmarking for generative models via k-wise human preferences. *arXiv preprint arXiv:2408.14468*, 2024. 2

[21] Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, Yanfeng Wang, and Weidi Xie. Intelligent grimm - open-ended visual storytelling via latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6190–6200, 2024. 5

[22] Quanfeng Lu, Wenqi Shao, Zitao Liu, Fanqing Meng, Boxuan Li, Botong Chen, Siyuan Huang, Kaipeng Zhang, Yu Qiao, and Ping Luo. Gui odyssey: A comprehensive dataset for cross-app gui navigation on mobile devices. *arXiv preprint arXiv:2406.08451*, 2024. 5

[23] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 6

[24] OpenAI. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/, 2024. Accessed: 2024-05-26. 9, 19

[25] Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S Nassar, and Peter W J Staar. Doclaynet: A large human-annotated dataset for document-layout segmentation. page 3743–3751, 2022. 5

[26] Nils Schaetti. Sfgram: a dataset containing thousands of scienc-fiction books and novels. https://github.com/nschaetti/EchoTorch, 2018. 5

[27] Danqing Shi, Weiwei Cui, Danqing Huang, Haidong Zhang, and Nan Cao. Reverse-engineering information presentations:

Recovering hierarchical grouping from layouts of visual elements. *Visual Intelligence*, 1(1):9, 2023. 5

[28] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5179–5187, 2015. 6

[29] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024. 9, 19

[30] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 19

[31] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 9, 19

[32] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. Caltech-ucsd birds. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 6

[33] Shiyao Wang, Qi Liu, Tiezheng Ge, Defu Lian, and Zhiqiang Zhang. A hybrid bandit model with visual priors for creative ranking in display advertising. In *Proceedings of the Web Conference 2021*, pages 2324–2334, 2021. 5

[34] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 9, 19

[35] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks 2021)*, 2021. 7

[36] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. NExT-GPT: Any-to-Any Multimodal LLM. In *Proceedings of the International Conference on Machine Learning*, 2024. 9, 19

[37] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Liang Pan Jiawei Ren, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, Dahua Lin, and Ziwei Liu. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 6

[38] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. VILA-U: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024. 9, 19

[39] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single

transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. 9, 19

[40] Zhe Xu, Dacheng Tao, Ya Zhang, Junjie Wu, and Ah Chung Tsoi. Architectural style classification using multinomial latent logistic regression. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 600–615. Springer, 2014. 5

[41] Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Yingcong Chen. Seed-story: Multimodal long story generation with large language model. *arXiv preprint arXiv:2407.08683*, 2024. 5

[42] Bingchen Zhao, Yongshuo Zong, Letian Zhang, and Timothy Hospedales. Benchmarking multi-image understanding in vision and language models: Perception, knowledge, reasoning, and multi-hop reasoning. *arXiv preprint*, 2024. 6

[43] Kaizhi Zheng, Xuehai He, and Xin Eric Wang. MiniGPT-5: Interleaved vision-and-language generation via generative vokens. *arXiv preprint arXiv:2310.02239*, 2023. 9, 19

[44] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Proceedings of the Advances in Neural Information Processing Systems*, 2023. 2

[45] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022. IEEE, 2019. 5

[46] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 6

[47] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7380–7399, 2021. 7

Figure 18. Examples of 23 Meta-Topics (**bold** font) and corresponding selected task (regular font) in Our OpenING Benchmark.

GUI Navigation: Cross-APP GUI Navigation

Please give the results of GUI navigation with interleaved image of GUI and text explanation. <BEGIN> image of GUI and text explanation. <BEGIN> Search for an Animation movie on Tubi: Movies & Live TV using Firefox and watch it in the movie app. <image>

Click the coordinate (620, 808). <image>Type the text \"animation movie on tubi\". <image> Click the coordinate (906, 900). <image> Click the coordinate (138, 372). <image> Click the Coordinate (836, 564). <image> Click the coordinate (527, 174). <image>

Interactive Visual Design: Product Design

<BEGIN> After the flood, clean drinking water is relatively difficult to find. Please efficiently use the \"brainstorming\" method to randomly design a product to solve the drinking water source problem, and present it to me using text and images.

Certainly, here is the translation: \n\n\"Alright, here is a design for you featuring a 'MAKE IT CLEAN,' a sustainable water purification and flood prevention sandbag. It uses high-density sponge materials internally to rapidly absorb floodwaters. Simultaneously, the contaminated floodwater is purified through a sewage purifier located at the bottom of the sandbag. Ultimately, the polluted floodwaters are transformed into a clean and sanitary source of drinking water.\" <image> The design effectively addresses the issues of traditional sandbags, such as slow water absorption, low drainage efficiency, and difficulty in transportation. It provides a sanitary source of drinking water for disaster-affected populations and improves the living environment conditions in post-disaster areas. <image>

Multimodal Layout Generation: Document with Layout Generation

<BEGIN> Please show the designed image of structured report and meet the following requirements: <BEGIN> Please provide a webpage displaying information on the \"Simon 2003 Annual Report\" with sections on easy navigation, search capability, and hyperlink functionality. Additionally, provide a layout description in JSON format and specify the coordinates of each element.

This is the designed webpage. The description of the layout is as follows: [{'bbox': [60.294117647058826, 204.01074558080813, 423.911879493464, 30.377246843434364], 'font': {'color': [0, 0, 0, 255], 'name': '/GTDJJJ+Futura-Bold', 'size': 1.0}, 'text': 'Simon 2003 Annual Report', 'type': 'Title'}, {'bbox': [60.294117647058826, 151.20772821969706, 281.00073529411765, 30.377259785353544], 'font': {'color': [0, 0, 0, 255], 'name': '/GTDJJJ+Futura-Bold', 'size': 1.0}, 'text': 'Investor Relations', 'type': 'Title'}, {'bbox': [60.294117647058826, 347.9766417297981, 172.3205882352941, 20.251489267676675], 'font': {'color': [0, 0, 0, 255], 'name': '/GTDJJJ+Futura-Bold', 'size': 1.0}, 'text': 'Easy Navigation', 'type': 'Section-header'}, {'bbox': [60.294117647058826, 597.1085473484849, 184.017766380718954, 20.25152809343433], 'font': {'color': [0, 0, 0, 255], 'name': '/GTDJJJ+Futura-Bold', 'size': 1.0}, 'text': 'Search Capability', 'type': 'Section-header'}, {'bbox': [60.294117647058826, 846.2405176767677, 194.28774509803918, 20.25151515151518], 'font': {'color': [0, 0, 0, 255], 'name': '/GTDJJJ+Futura-Bold', 'size': 1.0}, 'text': 'Hyperlink Enabled', 'type': 'Section-header'}, {'bbox': [60.294117647058826, 235.80643970959602, 468.84695833333336, 19.770075757575682], 'font': {'color': [0, 0, 0, 255], 'name': '/DVZRTP+Futura-Light', 'size': 1.0}, 'text': 'Thank you for downloading the interactive PDF version of ', 'type': 'Text'}, {'bbox': [60.294117647058826, 371.69655239898987, 411.7686609477124, 19.770075757575796], 'font': {'color': [0, 0, 0, 255], 'name': '/DVZRTP+Futura-Light', 'size': 1.0}, 'text': 'This PDF version of the annual report is very easy ', 'type': 'Text'}, {'bbox': [60.294117647058826, 620.8284968434343, 416.5318794934641, 19.77007575757574], 'font': {'color': [0, 0, 0, 255], 'name': '/DVZRTP+Futura-Light', 'size': 1.0}, 'text': 'It is easy to use the PDF document's built-in search ', 'type': 'Text'}, {'bbox': [60.294117647058826, 869.9604671717171, 898.0808321078432, 19.770075757575768], 'font': {'color': [0, 0, 0, 255], 'name': '/DVZRTP+Futura-Light', 'size': 1.0}, 'text': 'One of the best features of the PDF version of the annual report is its ability to contain hyperlinks. Hyperlinks ', 'type': 'Text'}, {'bbox': [192.68593137254902, 960.1811690656566, 639.6401793300654, 19.61477531565656], 'font': {'color': [0, 0, 0, 255], 'name': '/AXVZDV+Futura', 'size': 1.0}, 'text': 'Please click on a bookmark at left or the right arrow key at the top to begin!', 'type': 'Text'}] <image>

Interactive Image Editing: Interactive Portrait Image Editing

👤 Please show the revised image and corresponding explanations based on instructions: <BEGIN> Edit the man on the far left's attire. Swap out his blue checkered shirt for an orange plaid one, while adding sunglasses and a light blue cap to their look. Additionally, have him hold two yellow cocktails in his hands. <image>

🤖 The modified image shows ten individuals outdoors on a grassy landscape, with the surroundings remaining unchanged. The most noticeable alteration is the person on the far left, now dressed in an orange and white plaid shirt, wearing sunglasses and a light blue cap, and holding two yellow cocktails. <image>

Multimodal Exam: Geometric Problem Test

👤 Please answer the math problem with image and explanations: <BEGIN> As depicted in the figure, in trapezoid ABCD, AD is parallel to BC, and points E and F are the midpoints of AB and CD, respectively. We refer to line segment EF as the midline of trapezoid ABCD. What is the positional relationship between AD, EF, and BC? <image>

🤖 As shown in the figure, connect DE and extend it to intersect with the extension of CB at H. Since AD is parallel to BC, Therefore, ∠A = ∠ABH.\nThus, △DAE ≅ △HBE (by ASA), Consequently, DE = EH and AD = BH. Since DE = EH and DF = FC, Therefore, EF is parallel to BC, Thus, EF is parallel to AD and BC. <image>

Multimodal Information Summary: Interactive Multi-concept Image Composition

👤 <BEGIN> Here are several images of my blog, please summarize all the content in one image and explain why. <image>

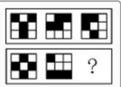🤖 Based on your description and the pictures you gave, I summarized the content as shown in the picture. Because the first picture expresses that dopamine has a reward function, such as enjoying delicious food, getting a good night's sleep, immersing in watching TV series, making in

The picture. Because the first picture expresses that dopamine has a reward function, such as enjoying delicious food, getting a good night's sleep, immersing in watching TV series, making your own cold drinks, taking a fragrant bath, and trying new things. The second picture expresses that serotonin has an analgesic function, such as jogging for 30 minutes, eating spicy hot pot, creative challenges, completing goals, climbing Mount Tai, and tasting dark chocolate. The third picture expresses that endorphins have a soothing effect. The fourth picture expresses that oxytocin has an intimate function, such as petting cats, praising others, camping with friends, listening to soothing music, lighting scented candles, and hugging a big doll. <image>

2D Reasoning: Multimodal Analogy & Biological Reasoning

👤 Please answer this question with image and text explanation: <BEGIN> What is the most appropriate answer to fill in the blank to make the sequence exhibit a certain regularity? <image>

🤖 Okay, based on the analogy relationship. When two shapes of different colors are added together, they become black blocks, while two shapes of the same color become black blocks, while two shapes of the same color become white. I generated this shape to present a certain pattern. <image>

👤 <BEGIN> May I ask what breed this cat is, and do you have more photos of this breed? <image>

🤖 This is a Munchkin cat. The Munchkin cat originated in the United States. It is gentle and enthusiastic by nature, likes to be with people, and can get along well with other animals. It enjoys lively environments and playing, and it is very willing to be held. The Munchkin cat has a short stature and is not skilled in climbing or jumping, making it suitable only for indoor life. However, it is extremely active and agile, able to adapt to fast running. Although both parents have short legs, sometimes long-legged offspring may be born. The Munchkin cat originated from a crossbreeding mutation that occurred in Louisiana, USA, in 1983. As breeders began to focus on purebred cats with non-purebred cats and studying the mutation, controversy over the experiment has grown along with this breed's rising popularity. In 1995, The International Cat Association (TICA) granted the Munchkin cat the title of \"new breed,\" being the only major registration organization to recognize the breed. Here are other photos of the Munchkin cat. <image>

Event Reasoning and Deductive Simulation: Dynamic Sports Event Analysis

<BEGIN> In the 21st round of the 2024 La Liga, Barcelona won 4-2 against Real Betis away. Please recreate the goals from that match using both images and text.

In the 21st minute, Torres scored with a close-range shot. <image> In the 48th minute, Yamal assisted Torres in scoring a goal. <image> In the 56th minute, Isco scored a goal for the home team. <image> In the 59th minute, Isco scored again to equalize the score. <image> In the 90th minute, Ferran assisted Félix in scoring a goal. <image> In the 92nd minute, Ferran Torres completed a hat-trick. <image>

Image-based 3D Reasoning: Interactive Novel View Synthesis

<BEGIN> This is a house. Its appearance features are:\"The main part of this house is mainly brown in color, there is a gate on the first floor with steps in front of the door with green grass on both sides, the whole consists of bricks, the roof is pitched, the roof is blue and green with chimneys, it does not have symmetry\". Can you guess what it looks like from the back? Please draw the picture and give descriptions. <image>

This is the back view. The walls are designed to resemble stacked logs, giving it a classic cabin-in-the-woods feel. The Roof features green tiles, adding a subtle contrast to the earthy tones of the walls. The overall design is simple, yet it evokes the charm of traditional, cozy cabins. The model's texture gives it a slightly handcrafted appearance, emphasizing its homely and traditional aesthetic. <image>

Cooking Tasks: Multimodal Recipe Generation

<BEGIN> How to make such an okra-baked eggs? Please teach me the steps in a pictorial format. <image>

Prepare the ingredients: 2 okras, 2 tomatoes, 2 eggs. <image> Cut the tomato at 1/5 from the top, and use a spoon to scoop out the flesh inside the tomato. <image> Wash the okra and slice it thinly. <image> Place the eggs and okra inside the hollowed-out tomatoes, then sprinkle a little salt and add some olive oil. <image> Put the cut-off top of the tomato back on. <image> Wrap the tomatoes well with aluminum foil; PS: Wrapping with aluminum foil is to prevent the loss of nutritional moisture, but it is not necessary. <image> Preheat the oven to 190 degrees Celsius for 25 minutes. PS: Each oven is different, so the temperature and time may vary slightly. <image>

Multimodal Information Recommendation: Interactive Film and Television Recommendation

Please output recommendations in the form of the poster and the corresponding introduction: <BEGIN> Do you have any must-watch animated movie recommendations?

Let me recommend to you three must-watch animated films!\n\n\"Spirited Away\"\nSynopsis: Chihiro and her parents are driving to their new home when they accidentally enter a mysterious tunnel on a countryside road—they find themselves in a strange world, a medieval town. Here, Chihiro meets a boy named Haku. One day, Chihiro discovers Haku is injured by a group of flying white paper spirits. To save the injured Haku, Chihiro embarks on her adventurous journey. <image> Zootopia\n\nSynopsis: The story takes place in a wonderful world where all mammalian species coexist harmoniously. Through her intelligence and hard work, the rabbit Judy successfully graduates from the police academy and joins the Zootopia Police Department. One day, while on duty, Judy encounters a fox named Nick. The two initially clash but end up partnering together to uncover a sensational secret hidden within Zootopia. <image> \n\n**Coco**\ n\n **Synopsis:** In an unexpected turn of events, Miguel finds himself transported to the Land of the Dead. Before the sun rises, he must receive a blessing from a family member; otherwise, he will be trapped in this world forever. Miguel decides to seek out the late musical legend, Ernesto de la Cruz, who might be his grandfather. Along the way, Miguel encounters the down-on-his-luck musician Hector and gradually uncovers long-hidden secrets about Ernesto de la Cruz. <image>

Multimodal Time Series Forecasting: Visual Traffic Forecasting

👤 <BEGIN> What will the traffic situation be like near the S2 Hulu Expressway in Pudong New Area, Shanghai, in one hour? Please provide an explanation and present it in the form of a picture. <image>

🤖 In one hour, certain sections of the S2 Hu-Lu Highway in the Pudong New Area of Shanghai will experience congestion. The road conditions near the area displayed on the entire map will be shown in the diagram. <image>

Geographical Tasks: Urban Planning and Development Simulation

👤 Please output the scheme in the form of both the image and the text explanation to meet the requirements: <BEGIN> This is the architectural planning map of the new residential area. How should the lake in the red frame be designed? What primary buildings should be around the lake? Please generate an aerial view to illustrate. <image>

🤖 As shown in the picture, the city harmoniously integrates with green water, with the river and three lakes connected by a green network. This networked ecological space system focuses on leisure and vitality, possessing unique ecological charm. <image> Dominant functions, productive service industries, headquarters office, hotel conferencing, technology research and development, cultural creativity, business leisure. Key projects, Cultural Expo Pavilion, Baishui New World, Zero-Carbon Exhibition Hall, and High-Line Sightseeing Corridor. <image>

Fashion Tasks: Interactive Virtual Try-on

👤 <BEGIN> Please generate the wearing effect of the clothing based on the photos of the clothes and the model that I provided, and provide an evaluation of the fitting effect. <image>

🤖 Certainly, here is the English translation:\n\n\"Alright, based on the image you provided, we have generated a render of how the clothes will look on you. The result after changing outfits is not satisfactory. The pattern on the chest of the white top is too vivid, which conflicts with the style of the pants. In summary, the clothes you had before the change are more suitable for your style.\" <image>

Multimodal Brainstorming: Academic Scenarios Brainstorming

👤 <BEGIN> Please draw a framework for nondestructive yield data measurement and tell me how to design such framework.

🤖 This image depicts a sophisticated framework for nondestructive yield data measurements using drone and smartphone technology integrated with an object detection model. Here's a conceptual breakdown to design this framework:\n\nData Acquisition:\nDrones equipped with cameras: These drones fly over the crop fields and capture high-resolution images. These images should cover various growth stages of the crop.\nSmartphones: Used by individuals on the ground to take close-up photos of the crops for more detailed analysis.\nData Processing:\nInput Images: The images collected by both drones and smartphones are fed into an object detection model.\nObject Detection Model (e.g., CSPRepResNet with PAN): This model is designed to identify and quantify various features such as the number of spikes, seed numbers per spike, and the seed size from the images. The model processes different layers (C3, C4, C5) and integrates information using a Path Aggregation Network (PAN) to enhance feature detection across scales.\nFeature Extraction:\nExtract specific features from the crops, such as the number of spikes per square meter and seed size, which are crucial for yield estimation.\nData Analysis:\nQuantitative Analysis: Using the detected features (e.g., seed numbers per spike, number of spikes per square meter, seed size) to estimate the crop yield.\nCrop Variety Identification: Identify the variety of the crop based on physical features detected in the imagery.\nOutput:\nCrop Yield Estimation: Calculate estimated yield based on the analyzed data.\nVisualization: Provide a user-friendly interface (possibly on the smartphone app) to display the analyzed data and yield predictions.\nFeedback Loop:\nAdjustments and Calibration: Adjust the model based on ground-truth data collected manually or feedback from the users to improve accuracy over time.\nThis framework would be beneficial for agricultural scientists, farmers, and agritech companies to monitor crop health and predict yields nondestructively and efficiently. <image>

Embodied AI Tasks: Autonomous Driving and In-door Navigation

<BEGIN> Picture yourself as an AI entity that views its surroundings through a camera. You've received three images taken at different moments in the past and present. Based on these, please make a prediction about the appearance of the upcoming frame. <image>

I walked down a wooden hallway, with a table on the left and a glass display on the right. Nearing the bar counter, I glanced left at the dining space. In the next frame, I'll move past the bar for a full view of the dining hall. <image>

Graph Generation: Mind Map Generation

<BEGIN> Summer has arrived. Have you been bitten by mosquitoes? Do mosquitoes prefer to bite people with type O blood? Please use a mind map to create a popular science presentation about mosquitoes. Show the image of map and the text explanation.

Impressions of mosquitoes <image> The structure of mosquitoes <image> The diet of mosquitoes <image> Three most common types of mosquitoes that bite in China <image> The steps of a mosquito sucking blood <image> The flying method of mosquitoes <image> The dangers of mosquitoes <image> The differences and similarities between mosquitoes and flies <image> Knowledge of mosquitoes <image>

Storybook Creation: Story Writing

<BEGIN> Please create a storybook named the night before Thanksgiving. The story is about a family celebrating the night before Thanksgiving together. Each part of this storybook should have a paragraph with a corresponding image.

A mother bakes three types of pie for Thanksgiving: pecan, pumpkin, and apple. The family is excited for the feast. <image> A boy carries a table and chairs down the stairs. <image>The family counted and polished our best silverware, then set the two tables with patience and care. <image> Children play in a room, calling out to each other, while their mother prepares a meal in the background, with a turkey. <image> Dad tripped over a hound while sliding a 30-pound turkey out of the oven. The turkey flew over the string beans and landed on a platter, where my brother and I caught it. Grandpa told us to hurry up and start carving, as the guests were starving. <image> The whole family celebrates togetherness and the end of a Thanksgiving feast around a roasted turkey. <image>

Multimodal Content Completion: Multimodal Report Completion

<BEGIN> Please complete the illustrated continuation of this educational piece on marine life, focusing on sea slugs: \n\nDiscover a fascinating animal every day — the Leaf Sheep Sea Slug \n\nThe Leaf Sheep Sea Slug (scientific name: Costasiella kuroshimae) \n\nAlso known as the Leaf Sheep, this charming marine creature boasts bright eyes and adorable antennae, giving it the appearance of a cartoon-like sheep. Its back is covered by only a thin shell, and it is an animal capable of photosynthesis. <image>

The leaf sheep can only grow up to five millimeters in length and has an adorable appearance. Its body is soft, with fluffy antennae, small and bright eyes, and a cute outer layer, resembling little sheep on land. Hence, it is named \"leaf sheep.\" <image> The leaf sheep is commonly found in the waters of Japan, Indonesia, and the Philippines. It feeds on grasses (algae). The leaf sheep utilizes the chlorophyll ingested into its body to produce nutrients, possessing kleptoplasts that can transform the chlorophyll from its food into its own body for its use. <image> Like most members of the sea slug family, Elysia marginata undergo external fertilization. After hatching and before feeding on thin algae, their bodies appear transparent. It is only after they begin consuming algae that their body color slowly starts to change. When Elysia marginata begins turning green, it signifies that they have reached maturity. The lifespan of the green leaf sea slug is only one year, but Elysia marginata can survive longer. <image>

Social Media Tasks: Plog and Social Media Content Generation

👤 <BEGIN> Traveling to Weihai in winter, I had planned to document this trip, but unfortunately, I forgot to save the photos. Could you help me create a visually rich plog for my use, highlighting the winter sea and snow of Weihai?

🐝 Weihai, a treasure city in Shandong, offers the splendid experience of watching snow by the sea and viewing the sea amidst the snow. It is truly breathtaking. <image> First, I checked in at Torch Street Eight! The houses only get sunlight from noon to dusk, which is very pleasant. <image> Then, I arrived at the seaside beneath Torch Eight Street! The waves were rolling layer upon layer, giving a sense of boundless <image> The next day, I went to Cat's Head Mountain. What sets this place apart from others is to pay attention to the one-way roads! Once you go, you can't come back the same way. <image> On the third day, we went to the long-anticipated Yuanyao Pier and the Watchtower chapel for sea weddings! From dusk to night! Two different experiences. <image>

Educational Tasks: Interactive Tutorial Generation

👤 Please show me the steps of the tutorial with interleaved images and text: <BEGIN> Please tell me how to add my engraving on the wood.

🐝 Place the template in the center of the top of the wood. Retrieve the template you made or printed, and place it on the wood. Align the template with the guidelines you drew. Add tape to the corners of the template to keep it in place while drawing. <image> Trace the outline of each word on the template with a pencil. Press the pencil tip into the paper and trace along the lines of the template. Focus only on the outline of each letter. Apply strong pressure so the strokes show up on the wood. Use a pencil so you can correct any mistakes. <image> Remove the template from the wood and check the visibility of the outline. If your outline was successfully traced, you will notice the pencil marks. If you cannot see the lines clearly, place the template back on the wood and retrace the letters' outline. You can erase any mistakes using a pencil eraser. <image> Use a permanent marker or paint pen to carefully write each letter. Mistakes made with a permanent marker or paint pen cannot be erased! Pay attention to the lines when tracing. Focus on drawing the correct position and shape of each letter. <image> Use different stroke widths and curl designs to embellish your letters. Once the letters are properly outlined, focus on making them look attractive. Apply broader strokes at the loops of the letters to enhance visual appeal. Add curls at the ends of the letters to increase their ornate appearance. Follow your template to remember where to add extra designs to your letters. <image> Use a pen or marker to go over the words to darken them. Adding a second layer to the letters with a paint pen will sharpen them and create a cleaner design. Ensure the first layer is dry before applying another one! Use quick-drying paint markers to save time and create bolder lines; quick-drying markers take only a few minutes to dry. <image>

Healthcare Tasks: Fitness and Health Consulting

👤 Please give answers in interleaved images and texts: <BEGIN> Please tell me how to perform a basic sit-up.

🐝 Bend your knees, and place the soles and heels of your feet flat on the ground. <image> Place both hands on the opposite shoulders, crossing your arms in front of your chest or behind your head. This provides you with a central point of elevation. <image> Pull the navel towards the spine and gently tighten the abdominal muscles. <image> Keep your heels on the ground and toes flat, first slowly and gently lift your head, followed by the shoulder blades. Focus your eyes on your bent knees while gently contracting your abdominal muscles. Lift off the floor until you reach a 90-degree angle, or when your elbows are above or at your knees. <image> Hold this position for one second. Slowly lower the torso back towards the floor, but try to keep it slightly elevated. This means not resting your back flat on the ground, but maintaining a slight yet relaxed arch. <image> Repeat steps 3-5 for the remaining part of the exercise. If you are a beginner, just do it two to three times and then gradually increase the number as your strength improves over time. Hopefully, you can also lose weight! <image>

Please show results in interleaved images and texts. <BEGIN> Please provide me with a few attractions and travel tips for Austria.

Famous Attractions and Routes: \nD1. Vienna State Opera: The music temple of Vienna. \nD2. Schönbrunn Palace: A royal residence in the suburbs of Vienna. \nD3. Salzburg: The hometown of Mozart, featuring many historical buildings and museums. <image> D4. Hallstatt Lake: Known as the \"most beautiful small town in the world.\"\n\nD5. Graz Old Town: A major city in Austria, it has preserved the complete appearance of the medieval era.\n\nD6. St. Wolfgang Lake District: Renowned for its beautiful lakes and mountains, as well as its fairytale-like castles. <image> D7. Wachau Valley: One of Europe's most beautiful valleys, offering stunning views of the Danube

River \nD8. Alpine Ski Resorts: Such as Kitzbühel and Lech, a paradise for skiing enthusiasts. \nD9. Innsbruck: Situated at the foot of the Alps, it features the famous Golden Roof and the Swarovski Crystal Museum. <image> Best Time:\n\nSpring (March-May): A season suitable for hiking and cycling all year round.\n\nSummer (June-August): The ideal season for swimming, boating, fishing, and kayaking.\n\nAutumn (September-November): The best season for enjoying gourmet food and vineyard scenery.\n\nWinter (December-February): The best season for skiing and other winter sports. <image> Transportation Guide: \n\nTrain: \n1. Salzburg -> Attnang-Puchheim -> Steeg-Gosau, then transfer to bus route 542. \n2. Salzburg -> Attnang-Puchheim -> Hallstatt, then take a boat to the town.\n\nBus: Take bus route 150 from Salzburg.\n\nDriving: Self-drive. <image>

Figure 19. Illustration of 23 pairwise comparison cases. The meta-topic name (**bold** font) and task name (regular font) are given for each case. Using the eight criteria detailed in Sec. C.3, human judges evaluate and compare the output of each model pair. Gold medal 🥇 is awarded to a model that generates content of significantly higher quality. In a tie setting, where the quality of outputs from the A model and B model is similar, Silver medal 🥈 is awarded to a model that generates relatively more favorable content. Errors that occur during model generation are highlighted by a red checkmark ❌ and Error Types . Models that pass all human moderation checks are marked with a green checkmark ✅.
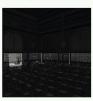
<BEGIN> How can the setting of a futuristic mail delivery system be portrayed in a new science fiction novel titled With the Night Mail: A Story of 2000 A.D.? Please directly write an example with illustrated images.

Model A (Anole)

With the Night Mail is a futuristic short story written by H.G. Wells in 1906. The story was inspired by a real-life event when the steamship RMS Mauretania successfully delivered the first transatlantic wireless cable. <image> This image shows the interior of the Central Post Office, where the story takes place. <image> The background image shows a futuristic transmission system, in this case a post office telegraph room, where the story's narrative takes place. <image>

❌

 Factual Error User asked for a new fiction, model generates an existing fiction

🥈 Model B (Human)

In the dim glow of the terminal's lights, the lower stages of the Great Postal Office outward mail towers buzz with activity. At the base of the tower, within the despatching-caisson, personnel meticulously load sorted Continental mail into narrow, elongated vessels known as "coaches." <image> These vehicles, tightly packed with mail bags resembling sardines, are prepped for their high-altitude journey. Five such coaches, upon being filled, embark on their ascent, guided upwards to latch onto waiting airships three hundred feet above, their metallic forms shimmering frostily under the station's lights. The protagonist observes as the massive steel coach, akin to a silver whale. <image>

✅

Model passes human moderation for all eight criteria

Multimodal Exam: Circuit Problem Test

👤 Please answer the physics question with image and explanations: <BEGIN> The elderly man's house needs to install a doorbell. However, since he has hearing difficulties, there should also be a flashing light bulb to alert him. Please design a circuit to meet his needs. <image>

🥇 Model A (Gemini1.5+Flux)

The circuit should include a power source, a switch, a doorbell, and a light bulb. When the switch is pressed, the circuit is closed, and both the doorbell and the light bulb are powered. The doorbell uses an electromagnet to move a hammer that strikes the bell. The light bulb uses a filament that glows when heated by an electric current. <image>

❌

| Poor Image Quality | The image is unrealistically rendered and fails to depict a well-rendered circuit design |

Model B (Next-GPT)

The image shows an elderly man sitting in his house, surrounded by various objects. There is a doorbell located on the wall, and it has a small flashing light bulb above it. The flashing light bulb is designed to alert the elderly man when someone is at the door, allowing him to open it safely. The circuit design for the doorbell and the flashing light bulb is simple, consisting of two wires. The first wire carries the signal from the doorbell, which is generated when the door is unlocked. This signal is sent to the flashing light bulb, which is located above the doorbell. The flashing light bulb is connected to a power source, such as a battery

❌

| No-Text or No-Image | No image generated |
| Factual Error | A doorbell typically generates a signal when a button is pressed, not when a door is unlocked. |

Storybook Creation: Fiction Writing

👤 <BEGIN> What are the must-see exhibitions and artifacts at the Musée du Louvre in France? Please share with me in a pictorial format. <image>

🥇 Model A (Emu3)

Certainly! The Musée du Louvre in France is renowned for its vast collection of art and artifacts from around the world. Some of the must-see exhibitions and artifacts include: 1.The Mona Lisa** - A famous painting by Leonardo da Vinci, housed in the Louvre's Great Hall … <image> 3.The Elgin Marbles** - A collection of ancient Greek artifacts, including the famous Elgin Marbles, which are on display in the Louvre's courtyard … <image> 5.The Terracotta Army** - A collection of terracotta soldiers from the First Emperor of China, displayed in the Louvre's courtyard… <image> 7.The Rosetta Stone** - Another ancient Egyptian artifact, this time with inscriptions in three scripts… <image>

❌

| Factual Error | Model make up fact. The Elgin Marbles, Terracotta Army, and Rosetta Stone are not in the Louvre. |
| Poor Image Quality | Unrealistic rendered image where elements placed in illogical positions |

Model B (MiniGPT-5MMD)

I'll give you a hint. The image should be a photograph of the Louvre Museum at night. <image> Give the image an artistic treatment. I'll give you a hint, it is also where the Mona Lisa painting hangs. I need an image of the Louvre at night in a style such as Impressionism or Realistic photography. Give it to me with caption and #hashtags for Twitter, LinkedIn & Instagram: <image> Ok, next step would be the Louvre Museum at night. Give it to me with caption and #hashtags for Twitter <image>

❌

Incoherent Content	Contents lack logical and natural flow
Image-Text Inconsistency	Text and image mismatched in style and scene.
Poor Text Quality	Repetitive and nonsensical mumbling
Incomplete Response	Model stops generating abruptly

Multimodal Layout Generation: Slide with Note Generation

<BEGIN> Please write a PowerPoint slides to explain the motivation of active learning. Write speaker notes for each slide.

🥇 Model A (Emu2)

Slide 23 <image>

❌

Factual Error	The content is irrelevant to active learning
Poor Image Quality	Rendered texts are unrecognizable, and graphics is too simple and vague
Poor Text Quality	No detailed description of the slide

Model B (Gemini1.5+Flux)

Slide 1: The Data Dilemma
Speaker Notes: Welcome, everyone. Today we're diving into the world of active learning. Let's face it, in a perfect world, we'd have access to endless perfectly labeled data. But the reality is, labeling data is expensive, time-consuming, and often, a real bottleneck in machine learning projects. *A sketch map located in the middle of the slide page shows a huge pile of unlabeled data with a frustrated stick figure holding a tiny label, trying to label them. On the other side of the map, there is a small pile of labeled data with a happy and relaxed stick figure.* ... <image>

✓

Model passes human moderation for all eight criteria

GUI Navigation: Cross-APP GUI Navigation

Please give the results of GUI navigation with interleaved image of GUI and text explanation. <BEGIN> Use Job Search by ZipRecruiter to find a software developer position, then use Google Keep to note down the name of the company. <image>

Model A (Emu2)

A screenshot of an iphone's home screen featuring several apps such as facebook and linkedin. <image>

❌

| Factual Error | Misinterpret the input and generate irrelevant answers and inaccurate image |
| Poor Image Quality | The rendered image is unrealistic, with misaligned icons of varying sizes |

🥇 Model B (Gemini1.5+Flux)

First, click on the ZipRecruiter app. Then, search for Software Developer positions. Look through the list of results, choose a company you like, and go to the company's website. <image> Then take note of the company name. Go back to the home screen by pressing the home button. Click the Keep Notes app. Create a new note by clicking the plus icon. Write down the company's name in the note. Finally, you can save the note. <image>

❌

| Incoherent Content | The generated images lacks coherence in style or entities with the input image |
| Image-Text Inconsistency | The images do not semantically align with the text, e.g., missing the list of companies and their websites. |

Interactive Image Editing: Interactive Landscape Image Editing

👤 Please give the result of edited image according to the input instruction and also give the description of editing results. <BEGIN> Lower the saturation of the entire image and increase the contrast. <image>

🥇 Model A (Emu2)

Lower the saturation and increase the contrast of the image. <image>

❌

| **Factual Error** | Inaccurate answer where there is no saturation in the input image |
| **Incoherent Content** | The generated image is incoherent with the input image |

Model B (GILL)

The image is a cityscape with a large number of people walking on the road.

❌

| **Factual Error** | Misinterpret the input and generate irrelevant answers |
| **No-Text or No-Image** | No image generated |

Multimodal Content Completion: Multimodal Report Completion

👤 <BEGIN> Please complete the copy by continuing with images and text for recommending oil-control lotion and emulsion products suitable for oily sensitive skin: Attention, oily sensitive skin! The magical combination of Streambean's oil-control lotion and emulsion has arrived! If you're uncertain about which lotion and emulsion to choose for oily sensitive skin, I suggest you try this one! Especially with summer approaching, the secretion of oil is particularly vigorous for us with oily skin, which can also lead to … <image>

🥇 Model A (VILA-U)

Streambean's oil-control lotion and emulsion is a great option for oily sensitive skin, as it provides a natural… <image> Streambean's oil-control lotion and emulsion are formulated with Trametes Versicolor, a plant that helps to reduce oil production, making it perfect… <image> The bottles are made of glass and have a minimalist design … <image> So, why not give Streambean's oil-control lotion and emulsion a try? <image> The Streambean's oil-control lotion and emulsion are designed to … <image>

❌

| **Incoherent Content** | Generated images are incoherent with the input image and across outputs, e.g., the appearance of product bottle. |
| **Poor Image Quality** | The rendered images are unrealistic, e.g., unbalanced scale of the product and model. |

Model B (Gemini1.5+Flux)

Firstly, let's talk about the lotion. <image>

❌

Incoherent Content	Generated images are incoherent with the input image and across outputs, e.g., the appearance of product bottle.
Poor Text Quality	Generated texts are too short to convey meaningful information
Incomplete Response	The model abruptly stops generating a textual response

Graph Generation: Figure Relationship Diagram Generation

<BEGIN> Could you provide a character relationship map for "Anna Karenina"? There are too many characters in this lengthy classic novel. Show the diagram and the text explanation.

🥈 Model A (SEED-X)

Certainly! Here is a character relationship map for "Anna Karenina":
| Character | Relationship to Anna Karenina || --- | --- | --- | | Levin | Levin is a close friend of Anna's and a potential suitor. He is in love with her, but she rejects him. || Vronsky | Vronsky is a military officer and Anna's lover. He is in love with Anna, but their relationship is complicated by societal expectations and Anna's husband, Karenin. <image>

❌

Poor Image Quality — The text in the image is rendered incorrectly and includes redundant figures

Incomplete Response — Model outputs incomplete and truncated response

Model B (Gemini1.5+FLux)

The main character is Anna Karenina. She has a love affair with Count Vronsky, a handsome and charming officer. The main character is Anna Karenina. She has a love affair with Count Vronsky, a handsome and charming officer. Anna is married to Alexei Karenin, a high-ranking government official. Their marriage is loveless, and Karenin is much older than Anna. Another important storyline involves Konstantin Levin, a landowner and friend of Stepan Oblonsky (Anna's brother). Levin is in love with Kitty, Stepan's sister-in-law. <image>

✅

Model passes human moderation for all eight criteria

Storybook Creation: Fiction Writing

<BEGIN> Hello, please help me generate an architectural design for a classical-style scenic area. Please show design ideas in interleaved images and texts.

Model A (SEED-LLaMA)

1. Step 1: Create a plan of the area with a central fountain and surrounding elements like trees, benches, and a gazebo. <image> 2. Step 2: Add decorative elements such as columns, statues, and a clock tower to the design. <image>

❌

Poor Image Quality — Rendered images are unrealistic

🥇 Model B (SEED-X)

1. Classical-style scenic area design ideas: a. Use classical architecture elements such as columns, arches, and pediments to create a sense of grandeur and elegance. b. Incorporate natural elements such as fountains, statues, and gardens to create a peaceful and relaxing atmosphere. c. Use lighting to highlight the architecture and create a dramatic effect. d. Incorporate seating areas and walking paths to encourage visitors to explore and enjoy the space. <image> 2. Design example: a. Classical-style building with columns, arches, and pediments. …. <image>

✅

Model passes human moderation for all eight criteria

2D Image Reasoning: Interactive Jigsaw Puzzle

👤 <BEGIN> These are some puzzle pieces. Please assemble them into a complete picture. Show the resulting image with the corresponding text explanation. <image>

Model A (GILL)

These are the four corners of the ancient city. <image>

Factual Error   Model produces an inaccurate image

🥇 Model B (Human)

Based on the puzzle pieces you provided, we have generated the complete image of the puzzle. Once fully assembled, the puzzle reveals a photograph of the Summer Palace. <image>

Model passes human moderation for all eight criteria

Event Reasoning and Deductive Simulation: Dream Analysis and Scene Reconstruction

👤 <BEGIN> I had a dream. Please help me visualize my dream into an image, and analyze why I had this dream, what implications and meanings it might have. Here is the content of my dream: I dreamed of two strange numbers, one on a digital clock in a room, and the other on the elevator display. Then, I started to flee.

Model A (Show-o)

The image of the image of the image of the image of the image of the image of the image of the image of the image of The image of the

image of the <image>

Factual Errors   Model fails to follow instruction and generates incorrect contents

🥇 Model B (Human)

Here is the image created based on the dream you described. The two strange numbers in this dream may symbolize certain unresolved mysteries in your life or things that cause you anxiety. The appearance of numbers in different places (a digital clock in a room and an elevator display) may indicate that these issues are pervasive, affecting your daily life. Your reaction of fleeing in the dream might represent your fear of these problems or a psychological desire to escape from them. <image>

Model passes human moderation for all eight criteria

Multimodal Information Summary: Interactive Multi-concept Image Composition

<BEGIN> These are four photos of sharing life. Please summarize all the content in one picture and write a paragraph under the title "A Breath of Life". <image>

Model A (GILL)

Tea cup and a book. <image>

❌

| Incoherent Content | The generated texts are to short |
| Poor Text Quality | The generated image is poorly laid out and lacks aesthetics. |

🥇 Model B (Emu2)

A person is reading a book and drinking tea. <image>

❌

| Poor Text Quality | The generated texts are to short |
| Poor Image Quality | The generated image lacks realism, where human hands are distorted |

Image-based 3D Reasoning: Interactive Novel View Synthesis

<BEGIN> This is a sofa. Its appearance features are: "This is a single-seat sofa, the whole is off-white, a little yellowish, the sofa is made of many folds, these folds are designed rather than due to prolonged use, the overall structure is axisymmetric". Can you guess what it looks like from the side? Please draw the picture and give descriptions. <image>

Model A (MiniGPT-5MMD)

These are the four corners of the ancient city. <image>

❌

Factual Error	User asked about sofas but received an irrelevant answer about a city
Incoherent Content	The generated image is incoherent with the input image
Poor Image Quality	The generated image lacks realism.

🥇 Model B (Human)

This is the side view. The upholstery appears to be smooth, possibly made from leather or a synthetic material with a glossy finish. The curved shape of the backrest and seat indicates a cozy and ergonomic design, which could offer good support for lounging or relaxation. The quilted pattern adds texture, giving the chair a stylish and comfortable look, perfect for a casual or contemporary living space. <image>

✅

Model passes human moderation for all eight criteria

Multimodal Information Recommendation: Interactive Film and Television Recommendation

Please output recommendations in the form of the poster and the corresponding introduction: <BEGIN> Could you recommend a few Brazilian TV series to me? <image>

🥇Model A (Gemini1.5+Flux)

Recommendation 1: 3%
This dystopian thriller series offers a thought-provoking plot and complex characters, making it a must-watch for fans of Black Mirror and The Hunger Games.
Recommendation 2:
Cidade Invisível (Invisible City) Dive into a world where Brazilian folklore meets urban fantasy in this visually stunning series. <image>

❌

| **Poor Image Quality** | The texts in the image are rendered inaccurately and illegible |

Model B (GILL)

I would recommend the following: <image>

❌

| **Poor Text Quality** | The generated texts are too short |
| **Incomplete Response** | Model stops generating textual response abruptly, and the response lacks details. |

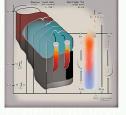**Multimodal Brainstorming:** Academic Scenarios Brainstorming

How do temperature and concentration gradients evolve over time in a star-finned adsorbent bed used in adsorption chillers? Please show an image for illustration. <image>

🥇Model A (SEED-X)

In a star-finned adsorbent bed used in adsorption chillers, temperature and concentration gradients evolve over time due to the adsorption and desorption processes. Here's an illustration to help visualize this:
1. Initial state: The adsorbent bed is initially filled with a cold, low-concentration gas (usually air or water vapor) and a warm, high-concentration liquid (usually water). The adsorbent bed is at a low temperature, and <image>
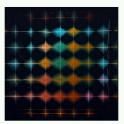

❌

| **Incomplete Response** | The textual response is incomplete |

Model B (Emu2)

I'm sorry, but I am unable to provide an image for illustration. <image>

❌

| **Factual Error** | Model fails to provide answers |

Multimodal Brainstorming: Academic Scenarios Brainstorming

In this task, you are given the first part of an event with both text and an image, and you need to complete the subsequent parts of the event by generating text and images that are natural continuation of the given first part. The input interleaved content is: <BEGIN> A wingsuit flyer is walking to his departure stage. <image> The wingsuit flyer puts his helmet on the ground. <image> The wingsuit flyer checks the condition of his equipment one by one because it's vital to life. <image>

Model A (GPT-4o+DALL-E3)

The wingsuit flyer carefully secures his helmet, preparing for the flight. <image> He approaches the edge of the platform, feeling the rush of anticipation. <image> With a final deep breath, he leaps off, soaring into the open sky. <image>

❌

Incoherent Content The generated images lack coherence with the input and are inconsistent across outputs.

🥇 Model B (Human)

The wingsuit flyer has finished checking, put on his helmet, and got ready. <image> The wingsuit flyer crouches slightly, lowers his weight, and attempts to take off in a perfect position. <image> The wingsuit glides in the sky above the mountains. <image>

✅

Model passes human moderation for all eight criteria

Multimodal Brainstorming: Academic Scenarios Brainstorming

<BEGIN> Please generate a panoramic view of the area based on my remote sensing satellite image. Also give interleaved text explanations for generated images. <image>

🥈 Model A (Gemini1.5+Flux)

This is a panoramic view of the area surrounding the buildings, with a focus on the roads and vegetation. <image>

❌

Factual Error Model produces inaccurate image

Model B (GILL)

The image is a satellite image of the area. <image>

❌

Factual Error Model produces inaccurate image

Poor Image Quality The generated image is blurry

Social Media Tasks: Chat with Memes

👤 You are a funny chatbot that responds to my small talk. Please output meme images to interact with me and chat with me. <BEGIN> Let's hurry up and get the tickets. <image>

🥇 Model A (GPT-4o+DALL-E3)

Oh, I'm feeling the excitement! Ready to pounce on those tickets. 😺 What's next on the agenda? <image>

✅
Model passes human moderation for all eight criteria

Model B (MiniGPT-5)

The image is a satellite image of the area. <image>

❌

Factual Error Model produces inaccurate image

Fashion Tasks: Fashion Trend Forecasting

👤 <BEGIN> What are the future autumn and winter color trends? Please explain to me with images and text.

🥇 Model A (Gemini1.5+Flux)

This fall and winter, earthy tones are trending. <image> Deep greens are especially popular, bringing a touch of nature to the colder months. <image> We're also seeing a lot of rich browns and tans, from camel to chocolate. <image>

✅
Model passes human moderation for all eight criteria

Model B (NExT-GPT)

Certainly! Let me present you with an image that captures the essence of future autumn and winter color trends. As you can see, the image showcases a variety of vibrant hues, including warm shades of red, orange, and yellow, as well as cooler tones like blue and purple. Each color is complemented by intricate patterns and textures, creating a visually captivating and trendy aesthetic. I hope this image inspires you to embrace the beauty of the upcoming autumn and winter season! <image>

❌

Offensive Content The distorted depiction of faces and hands of models may cause distress to readers

Poor Image Quality The generated image appears blurry

Cooking Tasks: Multimodal Cooking Assistant

👤 Please output instructions in interleaved images and texts: <BEGIN> What dishes can be made with the ingredients in the picture? Please provide a final product image. <image>

Model A (MiniGPT-5)

The dish can be made with vegetables and chili.. <image>

| Level | Levin is a Anna's and a potential love with her, but she Vronsky | Vronsky officer and Anna's with Anna, but relationship is

❌

Poor Text Quality — The textual response is too short and lacks detailed description

Poor Image Quality — The image does not resemble a completed dish and is lacking in aesthetic appeal

🏅 Model B (Gemini1.5+Flux)

You can make stir-fried mung bean sprouts with garlic chives using these ingredients. <image>

❌

Image-Text Inconsistency — The image depicts stir-fried noodles, which is inconsistent with the mung beans described in the text.

Education Tasks: Interactive Science Popularization

👤 <BEGIN> What is the magnetic field of a horseshoe magnet like? Please explain using diagrams and illustrations. <image>

Model A (GPT-4o+DALL-E3)

A horseshoe magnet's magnetic field is strongest between the two poles due to the curved shape that brings the poles closer together. The field lines emerge from the North pole, curve around in space, and enter the South pole. This concentration of field lines between the poles enhances its magnetic power, making it efficient for lifting and holding metal objects.
Below is an illustration showing the typical magnetic field lines:
![Horseshoe Magnet Field Lines](https://example.com/horseshoe-magnet-field-lines.jpg) These field lines illustrate how the magnetic force is directed from one pole to the other, concentrating in the space between the tips of the magnet. <image>
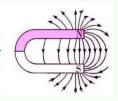
❌

Factual Error — The textual response is too short and lacks detailed description

🏅 Model B (Human)

The distribution of magnetic field lines of a horseshoe magnet starts from the N pole of the magnet, passes through the external space of the magnet, and points towards the S pole. Then, these magnetic field lines enter the interior of the magnet, proceeding from the S pole towards the N pole, and finally emanate from the N pole again, forming a closed loop. Outside the horseshoe magnet, the magnetic field lines exhibit a curved shape, while inside the magnet, the magnetic field lines display a straight shape. The tangent direction of these magnetic field lines indicates the direction of the magnetic field at that point, and the density of the magnetic field lines represents the magnitude of the magnetic induction strength. <image>

✅

Model passes human moderation for all eight criteria

Embodied-AI Tasks: Autonomous Driving and Indoor Navigation

<BEGIN> Picture yourself as an AI entity that views its surroundings through a camera. You've received three images taken at different moments in the past and present. Based on these, please make a prediction about the appearance of the upcoming frame. <image>

🥇 Model A (GILL)

The interior of the restaurant is a mix of modern and traditional. <image>

❌

Incoherent Content   The generated images are incoherent in style and content

Model B (MiniGPT-5MMD)

Here is the output.

❌

No-Text or No-Image   Model fails to generate image

Healthcare Tasks: Fitness and Health Consulting

Please give answers in interleaved images and texts: <BEGIN> Please tell me a method for practicing the leg movements of the breaststroke.

Model A (Gemini1.5+Flux)

One way to practice the breaststroke kick is to hold onto the side of the pool. <image>

Keep your body streamlined. <image>

Point your toes as you bring your feet together in the final motion. <image>

❌

Offensive Content   The distorted depiction of swimmer's body may cause distress to readers

Poor Text Quality   The textual response is too short and lacks description in details

🥇 Model B (GPT-4o+DALL-E3)

Method for Practicing Breaststroke Leg Movements
1. Kickboard Practice:
Use a kickboard to focus solely on your leg movement without worrying about arm strokes. <image>
2. Wall-Facing Kicks:
Hold onto the edge of the pool and practice your frog kick against the wall. Focus on the "whip" motion of the legs. <image>
3. Segments Practice:
Break down the kick into segments: bend, kick, snap, and glide to master each phase. Ensure smooth transitions between each. <image>

❌

Incoherent Content   The generated images are incoherent in style and content