

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

Pengfei Zhou^{*1}, Xiaopeng Peng^{*2}, Jiajun Song³, Chuanhao Li¹, Xu Zhao Pan¹, Yue Yang⁴¹, Ziyao Guo¹⁵, Hao Zhang¹, Yuqi Lin¹, Yefei He¹, Lirui Zhao¹, Shuo Liu¹, Tianhua Li¹⁴, Yuxuan Xie¹⁴, Xiaojun Chang⁶⁷, Yu Qiao¹, Wenqi Shao¹, Kaipeng Zhang^{†1}

Abstract

Multimodal Large Language Models (MLLMs) have made significant strides in visual understanding and generation tasks. However, generating interleaved image-text content remains a challenge, which requires integrated multimodal understanding and generation abilities. While the progress in unified models offers new solutions, existing benchmarks are insufficient for evaluating these methods due to data size and diversity limitations. To bridge this gap, we introduce GATE OpenING (OpenING), a comprehensive benchmark comprising 5,400 high-quality human-annotated instances across 56 real-world tasks. OpenING covers diverse daily scenarios such as travel guide, design, and brainstorming, offering a robust platform for challenging interleaved generation methods. In addition, we present IntJudge, a judge model for evaluating open-ended multimodal generation methods. Trained with a novel data pipeline, our IntJudge achieves an agreement rate of 82.42% with human judgments, outperforming GPT-based evaluators by 11.34%. Extensive experiments on OpenING reveal that current interleaved generation methods still have substantial room for improvement. Key findings on interleaved image-text generation are further presented to guide the development of next-generation models. The OpenING is open-sourced at <https://lancezpf.github.io/opening-benchmark.github.io/>.

1. Introduction

Building upon the remarkable understanding and generation capabilities of Large Language Models (LLMs) [1, 63, 64, 66], Multimodal LLMs (MLLMs) are making progress in various tasks [5, 42, 81, 84]. However, generating interleaved image-text content remains challenging [38, 62, 70], despite its important role in both research and applications (e.g., multimodal reasoning [11, 46], education [17, 37] and

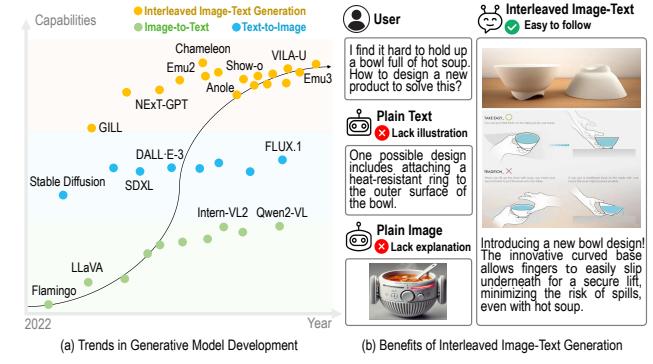


Figure 1. Motivation: (a) Rapid progress of interleaved image-text generation. (b) Interleaved content is essential to provide key information for complex real-world tasks (e.g., product design).

design [34, 58]). Since human brains can naturally combine visual and textual signals for more efficient information exchange [25, 31], achieving such integrated ability is crucial for advancing towards Artificial General Intelligence (AGI).

As shown in Fig. 1, the emergence of unified models that combine understanding and generation abilities opens new possibilities for interleaved generation [77, 92]. However, the lack of reliable benchmarks for evaluating interleaved image-text generation remains an obstacle [61, 70]. Most existing benchmarks evaluate text or image outputs separately and do not capture the complexities of generating both simultaneously [44, 60, 82, 83]. Interleaved benchmarks like OpenLEAF [4] and InterleavedBench [43] are limited in size, scope, and query diversity. For example, InterleavedBench includes only 815 instances across 10 tasks sourced from public datasets such as VIST [32], ActivityNet [36], and WikiHow [80], which is not representative of real-world needs and faces the risk of data contamination.

To fill the gap, we introduce OpenING, a comprehensive benchmark for evaluating open-ended interleaved generation. Unlike previous benchmarks, OpenING offers a broader set of real-world data and tasks (e.g., brainstorming, recommendations and content creation) derived from daily scenarios like fashion, cooking, and travel. As shown in Fig. 2 and Table 1, the curated OpenING includes 5,400 instances of

^{*}Equal contribution [†]Corresponding author ¹Shanghai Artificial Intelligence Laboratory ²Rochester Institute of Technology ³Renmin University of China ⁴Shanghai Jiao Tong University ⁵National University of Singapore ⁶University of Science and Technology of China ⁷MBZUAI



Figure 2. OpenING benchmark consists of 23 meta-topics (inner ring) which are further categorized into 56 specific tasks (see the number of tasks on the outer ring and details in Appendix Sec. A.1). Examples showcase interleaved generation in eight representative domains.

multi-step interleaved image-text content across 23 meta-topics and 56 tasks, with diverse, carefully designed queries for various topics. To tackle the challenges of gathering and standardizing data from disparate domains, we develop an efficient annotation pipeline and produced high-quality human-annotated data, reducing data contamination risks.

In addition, previous benchmarks often rely on GPT-based scoring metrics [4, 43], which are prone to be affected by the bias of GPT and potential data leakage in API uses [71]. To address the challenges of assessing open-ended multimodal generation, we introduce a robust judge model named IntJudge. We propose an Interleaved Arena to facilitate training data annotation and a Reference-Augmented Generation (RAG) approach to scale up the data size. Trained using the enhanced data pipeline, IntJudge achieves an 82.42% average agreement with human judgments, showing an 11.34% improvement over GPT-4o as a judge.

We evaluate representative interleaved generation methods using our OpenING. Key findings from our experiments include: 1) Generating coherent and high-quality interleaved content remains challenging for all models, as human-annotated content receives the highest rating compared to the generated content; 2) While integrated pipelines (e.g. Gemini+Flux) outperforms end-to-end models (e.g. Anole) with more coherent image-text content and higher quality, end-to-

end and two-stage generators (e.g. SEED-X) still hold great potential as the unified models continue to advance and can be combined with these methods; and 3) While text answers generated by GPT can be more informative than human-annotated answers, annotated natural images are still more favorable to generated images, highlighting the challenge of generating high-quality images. The major contributions of this paper are summarized as follows:

- **A High-quality Benchmark.** We present OpenING, a comprehensive benchmark for evaluating open-ended interleaved image-text generation. OpenING includes 5,400 human-annotated instances across 56 real-world tasks, aiming to challenge and improve interleaved generation methods and also support the development of judge models for assessing open-ended multimodal generation.
- **A Robust Judge.** We introduce IntJudge, a judge model for rating interleaved generation methods. We train IntJudge with an enhanced data pipeline, achieving an 82.42% agreement rate with human judgments and significantly outperforming GPT-based judge. Moreover, IntJudge has proven to be effective in assessing new unseen models.
- **A Comprehensive Leaderboard.** We provide detailed rankings and analysis of interleaved generation methods and compare evaluations from our IntJudge and GPT-4o

Benchmark	Data Coverage							Open-source	Offline Judge
	Meta-Topics	Tasks	Instances	Images	Steps	SpI			
OpenLEAF [4]	2	10	660	-	-	-	✗	✗	
InterleavedBench [43]	4	10	815	1,513	1,601	1.96	✓	✗	
OpenING (ours)	23	56	5,400	17,603	20,094	3.72	✓	✓	

Table 1. Comparison between OpenING and existing benchmarks. OpenING includes more comprehensive data and task coverage with an openly available judge model. Steps: a step is indicated by an input instruction or an output image-text pair; SpI: Steps per Instance.

with human judgments. Our studies suggest that although current open-source end-to-end models perform less favorably than integrated generation pipelines, two-stage generators based on unified models deserve future exploration for the development of interleaved generation.

2. Related Work

Interleaved Image-Text Generation.

Development of MLLMs has greatly pushed forward interleaved image-text generation [35]. Early models like Stable Diffusion [20, 53], DALL-E [52], and autoregressive (AR) methods such as VAR [65] and Lumina-mGPT [41] focused on unidirectional tasks, including image understanding and text-to-image generation. Flamingo [2] was the first to introduce capabilities for interleaved image-text content. More recent models, such as MiniGPT-5 [88] and SEED-series [23, 24, 79], achieve interleaved generation by combining the AR-based text generation and diffusion-based visual generation. Native AR models such as Emu-3 [70] and Chameleon [62] offer a unified framework capable of generating and reasoning over mixed-modal documents. Anole [16] reproduces the image generation capability of Chameleon through efficient fine-tuning on interleaved image-text data. Meanwhile, benchmarks for the evaluation of interleaved image-text generation are still emerging. Previous works, such as OpenLEAF [4] and InterleavedBench [43] focused on a small set of scenarios and lacked the depth and breadth for real-world applications. To achieve a more reliable and holistic evaluation of interleaved generation, we propose OpenING – a comprehensive benchmark consisting of real-world scenarios.

Evaluation of Open-ended Multimodal Generation. Evaluating open-ended multimodal generation is inherently challenging, due to the requirement of assessing both visual and textual quality in open domain [4, 73]. Existing text generation metrics, such as BLEU [49] and ROUGE [39], lack the measure of visual quality or text-image coherence. Visual quality metrics like FID [30] and IS [54], on the other hand, lack the consideration of textual elements. Metrics like CLIPScore [29] can measure text-image alignment but are unable to fully evaluate the quality of open-ended interleaved content, where multiple correct answers exist. GPT-based scoring [43, 86] provides better measurements to

assess the diversity and coherence of the interleaved outputs. However, GPT tends to be biased and favors the contents generated by its own [6, 71]. Human evaluation, though reliable, is not scalable due to its laborious nature. To bridge this gap, we introduce IntJudge, a judge model that is highly aligned with human judgments in evaluating the open-ended multimodal generation. To avoid the instability of subjective scores [14, 89], our IntJudge mainly evaluates models through pairwise comparisons in an arena manner.

3. OpenING Benchmark

3.1. Problem Definition

The task of interleaved image-text generation involves generating a sequence of text and images based on a given prompt. Each interleaved generation model (referred to as a multimodal agent) receives an input prompt \mathbf{P} , which can be text-only or include both texts and images. The multimodal agent outputs an interleaved image-text sequence: $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_N]$, where N is the number of steps. Each element $\mathbf{s}_i = <\mathbf{T}_i, \mathbf{I}_i>$ in step i consists of a text segment \mathbf{T}_i and an image \mathbf{I}_i . Each \mathbf{s}_i is generated based on the prompt \mathbf{P} and all outputs history as $\mathbf{s}_i = f(\mathbf{P}, \mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_{i-1})$, where f denotes the the generation function of an agent. The objective is to find an optimal output sequence set \mathbf{S}^* :

$$\mathbf{S}^* = \arg \max_{\mathbf{S}} \prod_{i=1}^N p(\mathbf{s}_i^* | \mathbf{P}, \mathbf{s}_1^*, \dots, \mathbf{s}_{i-1}^*), \quad (1)$$

where \mathbf{s}_i^* in each step is semantically consistent with the input prompt while the coherence throughout the entire sequence. The performance of an agent is evaluated based on how well the generated \mathbf{S} meets predefined criteria.

3.2. Data Curation

Collecting and annotating interleaved image-text data is inherently challenging due to the scarcity of high-quality data. Moreover, it is difficult to gather multimodal data from disparate domains and pair them consistently. We created OpenING over three months, with nearly 50 people involved in an efficient pipeline, which is shown in Fig. 3.

Topic Conceptualization. With the assistance of multiple AI agents, we brainstormed and identified the most relevant real-world scenarios that require interleaved image-text

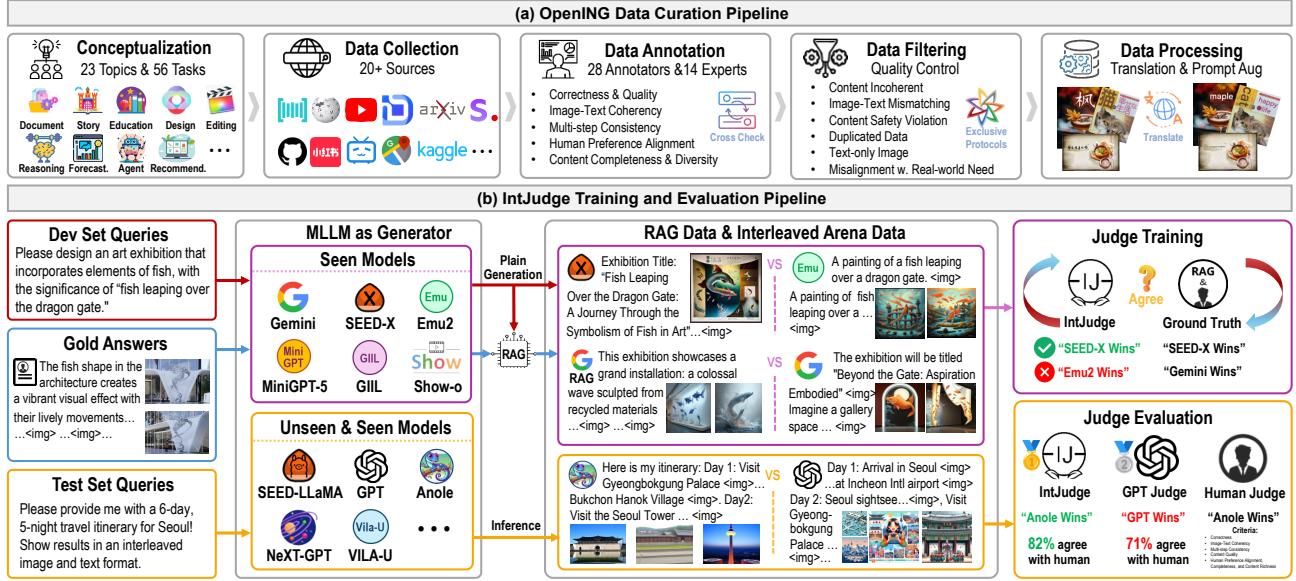


Figure 3. Overview of data curation and the proposed judge pipeline. (a) We construct our OpenING benchmark in a top-down manner, which involves five stages: conceptualization, data collection, annotation, filtering and processing. (b) We use the Dev Set of OpenING to train the proposed IntJudge and evaluate interleaved image-text generation on the Test Set to compare our IntJudge with human and GPT-4o.

generation. These insights were conceptualized into 23 meta-topics and divided into 56 specific tasks.

Data Collection and Annotation. We collected interleaved image-text data from more than 20 sources, including social media (e.g., REDnote¹) and YouTube², search engine (e.g. Google³), and open-source platforms (e.g. OpenDataLab [28]), etc. (see Appendix Sec. A.2). We designed an efficient annotation process involving a team of 28 professional annotators, monitored by 14 data experts throughout the process to ensure the highest quality. The maximum number of steps per instance was limited to ten to ensure usability with context restraints. All queries and answers are annotated manually via our developed tool, IntLabel.

Data Filtering and Quality Control. We performed cross-checks with annotators and data experts to ensure high standards of consistency, relevance, and coherence for each instance. Each task was required to include diverse sources and topics. In a few cases where data acquisition posed challenges, annotators were instructed to supplement the dataset with content generated by GPT-4o [48] and Stable Diffusion XL [51]. To further enhance data quality, we implemented a set of exclusive protocols for filtering unqualified data, which is illustrated in Fig. 3. The qualified data are resupplied to each task to ensure a certain amount after filtering.

Data Processing. We conduct the post-processing to ensure the linguistic consistency of the benchmark. We utilized the GPT-4o API to translate the annotated Chinese text to English in several common tasks, followed by a double-check

from data experts to verify the accuracy of translations. Additionally, for images containing Chinese characters, we implemented image translation⁴ from Chinese to English. We finally adjusted prompts in each task for desired generation results, which are detailed in Appendix Sec. A.3.

Dataset Splitting. As illustrated in Fig. 2, our OpenING benchmark ultimately includes 5,400 annotated instances, spanning 23 distinct meta-topics and 56 tasks. The annotated instances of OpenING are divided into a Dev Set (3,240 instances) and a Test Set (2,160 instances). The Dev Set supports the training of judge models, and the Test Set is used to evaluate the zero-shot performance of different models.

4. IntJudge Model

4.1. Interleaved Arena

Evaluating open-ended interleaved image-text generation is challenging due to 1) the complexity of assessing multiple images and text, and 2) the open-ended nature of interleaved generation, where multiple valid answers can exist. Given that pairwise comparison is more stable than subjective scoring [14], we develop the Interleaved Arena, which supports pairwise evaluation using three pipelines: human judges, GPT-based judges, and the proposed IntJudge.

In the Interleaved Arena, interleaved outputs from agents on the OpenING Test Set are saved in a unified format. In each evaluation round, judges compare outputs from two anonymous agents and rate the interleaved outputs based on seven criteria: Correctness, Image-Text Coherency, Multi-step Consistency, Content Quality, Human Preference Align-

¹<https://www.xiaohongshu.com/explore>

²<https://www.youtube.com>

³<https://www.google.com>

⁴<https://github.com/zyddnys/manga-image-translator>

ment, Completeness, and Content Richness (see Appendix Sec. B.1 for more details). To balance evaluation reliability and efficiency, we propose a roulette matching algorithm to sample E distinct battle pairs for each data instance:

Let \mathcal{K} represent the set of tasks, where each task $k \in \mathcal{K}$ contains D_k data instances. The set of agents in the arena is denoted as \mathcal{M} . For each task k , a permutation $\sigma_k \in A_{|\mathcal{M}|}$ is sampled by randomly shuffling the agent order, where $A_{|\mathcal{M}|}$ is the set of all permutations of agents in \mathcal{M} . The set of sampled battle pairs is given by:

$$\mathcal{P}_k = \{(\sigma_k(i \bmod |\mathcal{M}|), \sigma_k((i+1) \bmod |\mathcal{M}|))\}, \quad (2)$$

where $i = 1, 2, \dots, D_k$. Additional sampling rounds are conducted to obtain a total of E distinct battle pairs for each data instance, where $E \leq |\mathcal{M}|(|\mathcal{M}| - 1)/2$. To avoid duplication, we maintain a set $\mathcal{R}_{k,d}$ for the d -th round to track previously sampled pairs:

$$\mathcal{R}_{k,d} = \bigcup_{j=1}^{d-1} (\sigma_{k,j}(a), \sigma_{k,j}(b)). \quad (3)$$

For a newly sampled pair $\sigma_{k,d}(a)$ and $\sigma_{k,d}(b)$, we enforce:

$$(\sigma_{k,d}(a), \sigma_{k,d}(b)) \notin \mathcal{R}_{k,d} \text{ and } \sigma_{k,d}(a) \neq \sigma_{k,d}(b). \quad (4)$$

Under assumption of uniform distribution, we define the coverage time T_k to ensure all agents are evaluated in task k :

$$T_k = \left\lceil \frac{|\mathcal{M}|(|\mathcal{M}| - 1)}{2E} \cdot \frac{D_k}{|\mathcal{P}_k|} \right\rceil, \quad (5)$$

and the overall expected coverage time is given by:

$$E[T] = \frac{|\mathcal{M}|}{2} \cdot H_{|\mathcal{M}|} = \frac{|\mathcal{M}|}{2} \cdot \left(\sum_{i=1}^{|\mathcal{M}|} \frac{1}{i} \right), \quad (6)$$

where $H_{|\mathcal{M}|}$ is the $|\mathcal{M}|$ -th harmonic number.

4.2. Judge Pipelines

Human Judge. In the human judge, annotators compare outputs from two multimodal agents for each input prompt and select a winner based on seven predefined criteria. The voting results are used to rank interleaved generation methods based on their win rates. Since the previous studies [14, 89] noted that excessive ties cause inefficiency, our annotators are instructed to favor one agent in cases of a tie, denoting as Tie (A) or Tie (B) based on the slight preference.

GPT-based Judge. To enable scalability, we employ GPT-4o to automate the evaluation process. The GPT-4o is prompted to analyze interleaved outputs and decide the winner for each battle pair. Moreover, we use an additional prompt to obtain the score breakdown and explanations. While this allows for a scalable evaluation with explainability, GPT-based judges

still have a high error rate due to their prior bias and lack of alignment with human preferences. Besides, using GPT will bring concerns about privacy, data leakage, and cost.

IntJudge. To address issues in GPT-based evaluators, we propose IntJudge for more accurate evaluations and better alignment with human preferences. As an offline judge, IntJudge efficiently handles large-scale evaluations with consistent criteria, ensuring fair and reproducible results for benchmarking interleaved image-text generation. Specifically, we explored different MLLMs including InternLM-XComposer2.5 (InternLMX2.5) [85] and Qwen2-VL [68], ultimately choosing Qwen2-VL-7B as the foundational model for training IntJudge to achieve an optimal balance between efficiency and accuracy.

4.3. Training of IntJudge

To enhance the training of IntJudge, we propose a Reference-Augmented Generation (RAG) approach to scale up the training dataset. As illustrated in Fig. 3, we train our IntJudge model on the combination of human-annotated pairwise data from Dev Set and the RAG pairwise data. To generate scalable pairwise data using our RAG approach, models were provided with gold real-world answers from the Dev Set and prompted to generate responses based on these gold answers. A pairwise data is formed by pairing a plain generation result with an RAG-based output, with the RAG result assigned as the winner. A bag of models, including g seen interleaved generation methods are used for plain generation and RAG. The total training objective is defined as:

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

where $\lambda_1, \lambda_2, \lambda_3$ and λ_4 are weighting coefficients, \mathcal{L}_{CE} , \mathcal{L}_{CT} , \mathcal{L}_{MSE} , and \mathcal{L}_{PR} are respectively cross-entropy, contrastive, MSE, and pairwise ranking losses. The trained IntJudge was tested in a zero-shot setting on both unseen and seen models to validate its generalizability.

5. Experiments

5.1. Experimental Setup

Models. We test 10 representative methods and categorized them into three types: 1) **Integrated pipeline** consists of independent models for generating text and image in two stages, including GPT-4o+DALL-E-3 (DALL-E3) [8, 48] and Gemini1.5+Flux [9, 63]; 2) **Two-stage generator** uses a unified model architecture and outputs text and image in separate stages, including Emu2 [59], SEED-X [23], and Show-o [77]; 3) **End-to-end generator** outputs image-text results in a single stage, including GILL [35], NExT-GPT [74], MiniGPT-5 [88], SEED-LLaMA [22], and Anole [16]. We keep GPT-4o+DALL-E3, Anole, SEED-LLaMA, and NExT-GPT as unseen models for IntJudge validation, and the rest are models seen in IntJudge training.

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	83.28%	86.03%	68.17%	78.55%	82.49%	82.69%	82.03%	82.43%	87.46%	91.49%	75.49%	84.23%
GPT-4o+DALL-E3	78.42%	81.39%	65.21%	75.15%	85.70%	85.99%	85.58%	85.82%	85.02%	86.92%	72.22%	80.68%
Gemini1.5+Flux	65.57%	65.82%	49.31%	61.85%	71.75%	71.76%	71.12%	71.56%	68.30%	69.73%	54.47%	65.41%
SEED-X	51.98%	49.49%	34.70%	49.65%	54.82%	55.12%	54.11%	55.03%	49.86%	49.58%	33.57%	49.72%
Anole	51.90%	52.17%	36.46%	51.52%	53.36%	53.13%	52.58%	53.10%	53.42%	52.04%	33.92%	51.33%
SEED-LLaMA	44.30%	42.12%	29.11%	44.56%	40.96%	40.87%	40.46%	40.96%	50.13%	47.71%	31.57%	48.48%
Emu2	40.89%	37.07%	23.42%	41.84%	41.72%	41.63%	40.58%	41.85%	36.28%	33.79%	21.87%	39.51%
Show-o	36.28%	34.02%	21.63%	39.84%	30.77%	30.22%	29.61%	30.62%	31.49%	21.08%	12.48%	32.87%
NExT-GPT	33.67%	26.93%	17.09%	35.36%	22.61%	22.39%	22.11%	22.74%	30.96%	21.70%	13.36%	32.58%
MiniGPT-5	30.69%	26.72%	17.11%	35.09%	28.64%	28.37%	28.02%	28.64%	24.47%	15.46%	9.91%	27.85%
GILL	25.80%	19.57%	12.71%	30.23%	30.55%	30.24%	29.65%	30.62%	24.87%	19.72%	12.82%	30.32%

Table 2. 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 Tie (.5): Count a tie as 0 and 0.5 wins for a model in a battle, respectively.

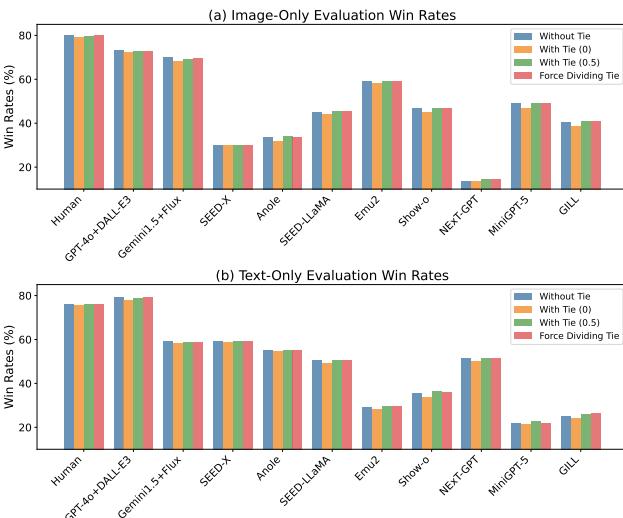


Figure 4. Model win rates under image-only and text-only settings across different models, ranked by human judgments.

Evaluation Metrics. We assess model performance using two key metrics: win rate and agreement. **Win rate** indicates how often a model wins in pairwise comparisons. We consider four methods for handling ties: 1) Force Dividing Tie (FDT): We force judges to assign ties with a more leaning model in rules and prompts, ensuring that every comparison results in a decisive outcome. If a tie leans towards model A ($\text{tie}(A)$), it counts as a win for A; similarly for B. This method allows for clear rankings without ambiguity. 2) Without Tie (w/o Tie): Tied comparisons are excluded; only matches with a clear winner are considered; 3) With Tie counted as 0 (w/ Tie (0)): Ties are included but do not contribute to the win count of either model; 4) With Tie counted as 0.5 (w/ Tie (.5)): Each tie contributes half a win to both models. **Agreement** measures the consistency between different evaluators (e.g., automated pipelines and human judgments) under the same tie-handling strategies. It

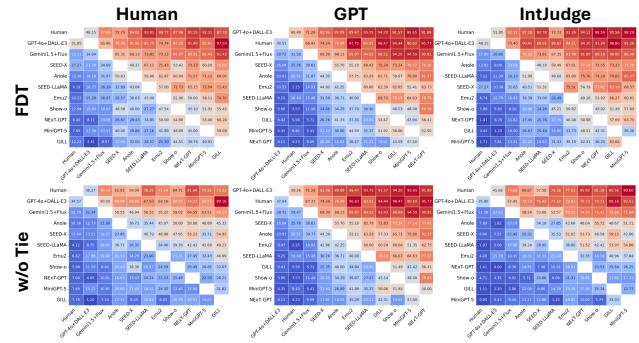


Figure 5. Win rate matrix of human and ten MLLM models, evaluated by human, GPT-4o, and our IntJudge, respectively.

reflects how often the evaluators concur in their assessments.

5.2. Overall Evaluation

Evaluation of Three Judges. We conduct experiments to evaluate the performance of different models using the win rate and agreement metrics. Table 2 showcases the win rates of various models under different judge methods, including Human, GPT-based, and IntJudge-based Evaluations. The sampling round E is set in 2 to form 4,320 battle pairs. It is found that the integrated pipelines like GPT-4o+DALL-E3 and Gemini 1.5+Flux consistently perform better across all evaluators, while the end-to-end models like MiniGPT-5, GILL, and NExT-GPT exhibited lower performance.

Pairwise Model Performance. We visualize the pairwise comparison results of all methods in Fig. 5, evaluated by human, GPT-4o, and IntJudge, respectively. The heat map reveals win-loss relationships, where warmer colors represent higher win rates and cooler colors vice versa. Notably, GPT-4o+DALL-E3 and Gemini 1.5+Flux demonstrate strong win rates against other models, even comparable to the manually annotated output under GPT evaluation.

Text-only and Image-only Evaluation. To explore whether text or image has a greater impact on the model performance, we evaluate models on text-only and image-only outputs on

Evaluator	FDT				w/ Tie				w/o Tie			
	Average	Seen	Unseen	HM	Average	Seen	Unseen	HM	Average	Seen	Unseen	HM
Random	49.83%	49.86%	49.79%	49.83%	32.60%	32.03%	33.18%	32.60%	50.00%	48.36%	51.89%	50.06%
GPT-4o	71.08%	73.33%	68.77%	70.98%	51.93%	54.95%	48.82%	51.70%	74.58%	77.54%	71.43%	74.36%
InternLMX2.5-7B	56.81%	55.73%	57.92%	56.81%	40.26%	40.19%	40.33%	40.26%	61.05%	61.21%	60.97%	61.09%
Qwen2-VL-7B	61.61%	61.59%	61.63%	61.61%	32.81%	31.16%	34.50%	32.75%	80.77%	81.15%	80.23%	80.69%
IntJudge-7B (Ours)	82.42%	84.05%	80.75%	82.37%	66.45%	69.02%	63.80%	66.31%	91.11%	92.38%	89.55%	90.94%

Table 3. Agreement rate between different MLLM-based judges and human judgments in different metrics. HM: Harmonic Mean.

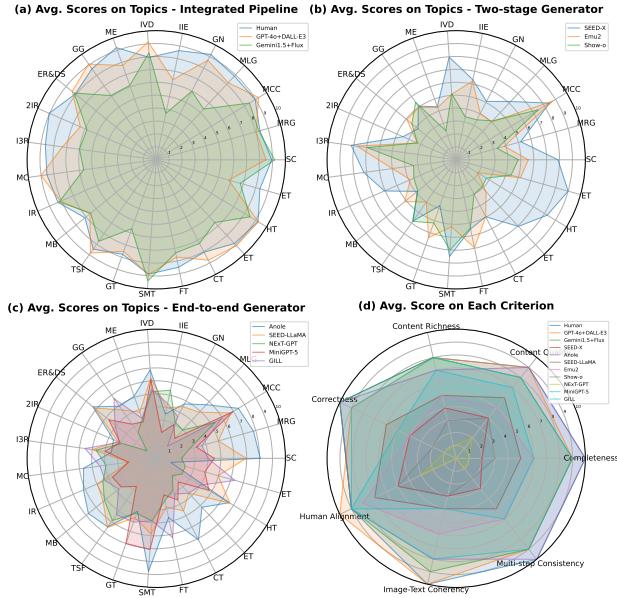


Figure 6. Evaluation results of GPT-based scores. (a)-(c): Average score of all criteria on each meta-topic for different kinds of models. (d) Average score of all meta-topics on each criterion.

the same sampled pairs. Fig. 4 shows that MiniGPT-5 and GILL do not perform well mainly due to the low quality of text outputs. SEED-X and NExT-GPT achieve higher win rates on text-only evaluation, however, the lower quality of generated images hinders their opportunity to achieve higher ranking in Table 2. It is observed that text generated by GPT-4o even outperforms real-world content annotated by human, highlighting the superior language capabilities of GPT-4o.

GPT-based Scoring. As shown in Fig. 6, detailed scores from GPT-based evaluations are provided to support explainable performance analysis of different models. It is observed that GPT-4o+DALL-E3 performs suboptimally in metatopics like Interactive Image Editing (IIE) and Embodied-AI Tasks (ET), possibly due to the lack of relevant training data in these scenarios. GPT-4o also showcases the inherent biases to its own generation results. For example, GPT-4o gives 10 scores to its own answers in Human Preference Alignment. In contrast, outputs annotated by humans only achieve an average score of 9 in human preference alignment.

Agreement with Human. Table 3 shows the agreement between different evaluators and human judgments. We im-

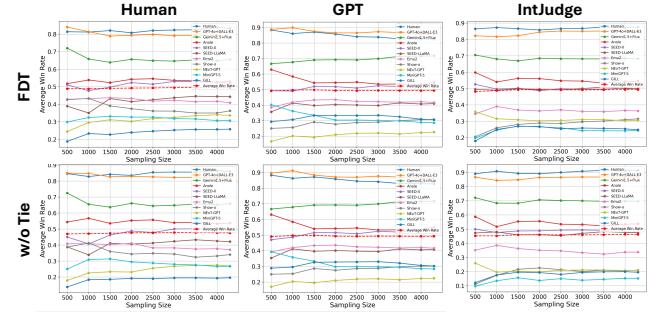


Figure 7. Effect of sampling size on evaluation reliability.

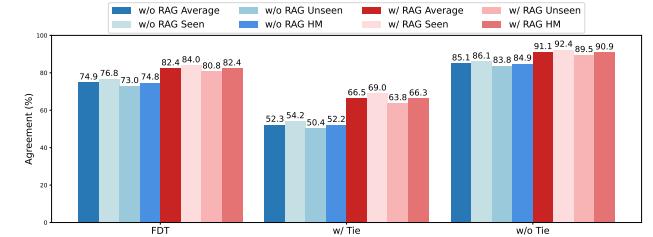


Figure 8. Comparison of agreement with human judgments for IntJudge trained without and with RAG data.

plement random guess (Random) as a baseline. The results indicate that IntJudge generally achieved higher agreement with human judgments (82.42% in FDT) compared to GPT-based evaluation (71.08% in FDT), suggesting its potential for scalable evaluation of interleaved image-text generation.

5.3. Ablation Studies

Ablation on Sampling Size. We evaluate the effect of sample size on evaluation stability and reliability. Fig. 7 illustrates the trend of win rates across varying sampling sizes. As the sample size increases, the win rates approach stability and show minimal variation across further increases. This stabilization suggests that our sampling number of 4,320 battle pairs is able to support the robust evaluation results.

Ablation on Judge Training Data. We investigate the influence of incorporating RAG data on the performance of the IntJudge. The comparison is conducted between two training configurations: one utilizing only the arena data (6,014 samples) and the other augmented with RAG data (25,982 samples). As illustrated in Fig. 8, with RAG data included, the FDT agreement on unseen models increases by 7.8%, demonstrating the effectiveness of our RAG-based strategy.

Ablation on Image Generator. We sample 200 data in-

Method	FDT	w/o Tie	w/ Tie (0)	w/ Tie (.5)
Human+Human	88.39%	92.23%	84.82%	88.84%
Human+Flux-dev	11.61%	7.77%	7.14%	11.16%
GPT+DALL-E3	49.51%	45.10%	22.33%	47.57%
GPT+Flux-dev	50.49%	54.90%	27.18%	52.43%
Gemini+Flux-sch	41.25%	41.43%	23.39%	42.14%
Gemini+Flux-dev	58.75%	58.57%	39.11%	57.86%
SEED-X+SEED-X	9.82%	5.15%	4.46%	11.16%
SEED-X+Flux-dev	90.18%	94.85%	82.14%	88.84%

Table 4. Evaluation results of interleaved content when basic text output combined with different image generation models.

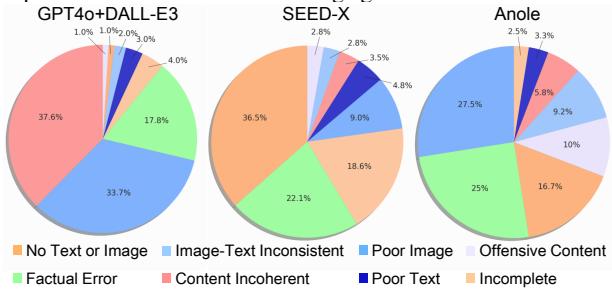


Figure 9. Error distribution of three models: GPT-4o+DALL-E3 (integrated), SEED-X (two-stage), and Anole (end-to-end).

stances across tasks to assess the influence of different image generators on interleaved performance. Table 4 presents a comparison of several basic text generation methods combined with different image generators. We assign the text generated by Gemini with Flux-schnell (Flux-sch) and Flux-dev to explore the difference. The results highlight that advanced image generators like Flux-dev are crucial for improving the overall quality of interleaved content. Meanwhile, it is noted that the generation efficiency of Flux-dev is significantly slow. This prompts further exploration of flow models to better balance generation quality and efficiency.

5.4. Analysis and Discussions

Error Analysis. We conduct an error analysis on 200 sampled instances where the output of models performed worse than human output. Fig. 9 shows the frequency of error types across three different types of models, providing insights into their specific performance limitations. GPT-4o+DALL-E3 suffers from content incoherency and inconsistency since it is hard for DALL-E3 to generate multiple images in the same style. Poor image quality is the main problem of Anole, as its finetuning data for image generation is insufficient. While most outputs by SEED-X have multiple errors, the inexistence of text or image content is still the major problem.

No-Image and No-Text Ratios. Table 5 presents the no-image, no-text, and no-image-and-text ratios of different models, representing the proportion of instances where the models fail to generate visual content, textual content, and both. Human, GPT-4o+DALL-E3, and Genimi1.5+Flux ex-

Method	Ratio		
	No-Image	No-Text	No-I&T
Human	0.00%	0.00%	0.00%
GPT-4o+DALL-E3	0.23%	0.00%	0.00%
Gemini1.5+Flux	0.09%	0.09%	0.09%
SEED-X	23.17%	4.64%	4.46%
Anole	19.46%	2.00%	1.30%
SEED-LLaMA	4.77%	0.05%	0.00%
Emu2	0.00%	15.10%	0.00%
Show-o	0.00%	7.74%	0.00%
NExT-GPT	43.97%	0.09%	0.09%
MiniGPT-5	0.27%	26.54%	0.00%
GILL	19.95%	13.43%	0.28%

Table 5. The ratios of No-Image, No-Text, and No-Image-and-Text (No-I&T) outputs relative to the total number of generated samples.

hibited near-zero failure rates (except that image generation on certain sensitive topics is prohibited according to content safety policies), indicating consistent multimodal generation. On the other hand, models like SEED-X and NExT-GPT showed a high no-image ratio. This can be attributed to poor instruction following and generation ability. These findings suggest that models need to generate both images and text properly to achieve higher rankings on OpenING.

Findings and Discussions. We discuss key findings from our experiments to inspire future works: **1)** While all generative models ranked lower than Human in interleaved generation, end-to-end models still lagged significantly behind integrated pipelines like GPT-4o+DALL-E3 in interleaved generation tasks. Two-stage generation methods based on unified models are also expected to be further improved. **2)** Natural images consistently outperform generated images, highlighting the challenge of generating high-quality images. **3)** Text generated by GPT can match or even surpass the quality of human-annotated text, showcasing the effectiveness of LLMs in producing rich and informative content. **4)** Foundational generative models play a big role. For example, when combining GPT, Gemini, and SEED-X with Flux-dev, respectively, all models demonstrate higher performance. **5)** As large-scale data is crucial for training judge models, using our RAG method to scale up data beyond manual annotation contributed to the building of a more robust judge model.

6. Conclusion

We introduce OpenING, a comprehensive benchmark designed to evaluate open-ended interleaved image-text generation. OpenING overcomes the limitations of existing benchmarks by including more diverse data and tasks grounded in real-world scenarios. To address the challenges in assessing open-ended multimodal generation, we propose a robust judge model IntJudge, which is trained on both human-annotated and RAG-based data from the Dev Set of OpenING. We evaluate various interleaved generation methods

on the Test Set of OpenING, revealing the significant challenges for generating coherent and high-quality interleaved image-text content. Ablation studies further demonstrate the effectiveness of our RAG-based data pipeline in training IntJudge. Looking forward, increasing the size and diversity of the data could further enhance the practical relevance of interleaved generation benchmarks. We anticipate that OpenING will inspire further research in MLLMs, and support the development of multimodal evaluation models.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. [1](#)
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Proceedings of the Advances in Neural Information Processing Systems*, 2022. [3](#)
- [3] Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. BoQ: A place is worth a bag of learnable queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17794–17803, 2024. [16](#)
- [4] Jie An, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Lijuan Wang, and Jiebo Luo. Openleaf: Open-domain interleaved image-text generation and evaluation. *arXiv preprint arXiv:2310.07749*, 2023. [1](#), [2](#), [3](#)
- [5] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. [1](#)
- [6] Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. Benchmarking foundation models with language-model-as-an-examiner. In *Proceedings of the Advances in Neural Information Processing Systems*, 2024. [3](#)
- [7] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. *CVPR*, 2022. [16](#)
- [8] 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. [5](#), [28](#), [29](#)
- [9] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024. Accessed: 2024-11-05. [5](#), [28](#), [29](#)
- [10] 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. [15](#)
- [11] Ju-Seung Byun, Jiyun Chun, Jihyung Kil, and Andrew Perault. ARES: Alternating reinforcement learning and supervised fine-tuning for enhanced multi-modal chain-of-thought reasoning through diverse AI feedback. *arXiv preprint arXiv:2407.00087*, 2024. [1](#)
- [12] 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. [16](#)
- [13] 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. [17](#)
- [14] 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. [3](#), [4](#), [5](#), [25](#)
- [15] 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. [15](#)
- [16] 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. [3](#), [5](#), [28](#), [29](#)
- [17] Daniel Claman, Emre Sezgin, et al. Artificial intelligence in dental education: Opportunities and challenges of large language models and multimodal foundation models. *JMIR Medical Education*, 10(1):e52346, 2024. [1](#)
- [18] 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. [16](#)
- [19] Alexey Dosovitskiy, German Ros, Felipe Codella, 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. [17](#)
- [20] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the International Conference on Machine Learning*, 2024. [3](#)
- [21] 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. [17](#)
- [22] 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. [5](#), [28](#), [29](#)
- [23] Yuying Ge, Sijie Zhao, Chen Li, Yixiao Ge, and Ying Shan. Seed-data-edit technical report: A hybrid dataset for instruc-

- tional image editing. *arXiv preprint arXiv:2405.04007*, 2024. 3, 5, 15, 28, 29
- [24] Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making LLaMA SEE and Draw with SEED Tokenizer. In *Proceedings of the International Conference on Learning Representations*, 2024. 3
- [25] Matthew F Glasser, Timothy S Coalson, Emma C Robinson, Carl D Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F Beckmann, Mark Jenkinson, et al. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016. 1
- [26] Yulia Gryaditskaya, Mark Sypesteyn, Jan Willem Hoftijzer, Sylvia Pont, Frédéric Durand, and Adrien Bousseau. Open-sketch: A richly-annotated dataset of product design sketches. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 38, 2019. 15
- [27] 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. 16
- [28] Conghui He, Wei Li, Zhenjiang Jin, Chao Xu, Bin Wang, and Dahua Lin. OpenDataLab: Empowering general artificial intelligence with open datasets. *arXiv preprint arXiv:2407.13773*, 2024. 4
- [29] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 3
- [30] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the Advances in Neural Information Processing Systems*, 2017. 3
- [31] Judith Holler and Stephen C Levinson. Multimodal language processing in human communication. *Trends in Cognitive Sciences*, 23(8):639–652, 2019. 1
- [32] Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016. 1
- [33] 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. 15
- [34] Hyung-Kwon Ko, Gwanmo Park, Hyeon Jeon, Jaemin Jo, Juho Kim, and Jinwook Seo. Large-scale text-to-image generation models for visual artists’ creative works. In *Proceedings of the International Conference on Intelligent User Interfaces*, 2023. 1
- [35] 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. 3, 5, 28, 29
- [36] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 1
- [37] Ehsan Latif, Gengchen Mai, Matthew Nyaaba, Xuansheng Wu, Ninghao Liu, Guoyu Lu, Sheng Li, Tianming Liu, and Xiaoming Zhai. Artificial general intelligence (AGI) for education. *arXiv preprint arXiv:2304.12479*, 1, 2023. 1
- [38] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekerman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. In *Proceedings of the Advances in Neural Information Processing Systems*, 2024. 1
- [39] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the ACL Workshop: Text Summarization Branches Out*, pages 74–81, 2004. 3
- [40] 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. 15
- [41] Dongyang Liu, Shitian Zhao, Le Zhuo, Weifeng Lin, Yu Qiao, Hongsheng Li, and Peng Gao. Lumina-mGPT: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. *arXiv preprint arXiv:2408.02657*, 2024. 3
- [42] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proceedings of the Advances in Neural Information Processing Systems*, 2024. 1
- [43] Minqian Liu, Zhiyang Xu, Zihao Lin, Trevor Ashby, Joy Rimchala, Jiaxin Zhang, and Lifu Huang. Holistic evaluation for interleaved text-and-image generation. *arXiv preprint arXiv:2406.14643*, 2024. 1, 2, 3
- [44] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. MMBench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 1
- [45] 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. 15
- [46] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhui Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. In *Proceedings of the Advances in Neural Information Processing Systems*, 2024. 1
- [47] 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. 16
- [48] OpenAI. Hello GPT-4o. <https://openai.com/index/hello-gpt-4o/>, 2024. Accessed: 2024-05-26. 4, 5, 28, 29
- [49] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine

- translation. In *Proceedings of the annual meeting of the Association for Computational Linguistics*, 2002. 3
- [50] 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. 15
- [51] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *Proceedings of the International Conference on Learning Representations*, 2024. 4
- [52] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3
- [53] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3
- [54] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Proceedings of the Advances in Neural Information Processing Systems*, 2016. 3
- [55] Nils Schaetti. Sfgram: a dataset containing thousands of scienc-fiction books and novels. <https://github.com/nschaetti/EchoTorch>, 2018. 15
- [56] 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. 15
- [57] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsom: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5179–5187, 2015. 16
- [58] Francesco Stella, Cosimo Della Santina, and Josie Hughes. How can LLMs transform the robotic design process? *Nature Machine Intelligence*, pages 1–4, 2023. 1
- [59] 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. 5, 28, 29
- [60] Zhiyu Tan, Xiaomeng Yang, Luozheng Qin, Mengping Yang, Cheng Zhang, and Hao Li. Evalalign: Supervised fine-tuning multimodal llms with human-aligned data for evaluating text-to-image models. *CoRR*, 2024. 1
- [61] Zineng Tang, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, and Mohit Bansal. CoDi-2: In-context interleaved and interactive any-to-any generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27425–27434, 2024. 1
- [62] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 1, 3, 29
- [63] 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. 1, 5, 28, 29
- [64] InternLM Team. InternLM: A multilingual language model with progressively enhanced capabilities, 2023. 1
- [65] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In *Proceedings of the Advances in Neural Information Processing Systems*, 2024. 3
- [66] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambo, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1
- [67] 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. 16
- [68] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 5
- [69] 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. 15
- [70] 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. 1, 3, 28, 29
- [71] Yidong Wang, Zhuohao Yu, Wenjin Yao, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, et al. PandaLM: An automatic evaluation benchmark for llm instruction tuning optimization. In *Proceedings of the International Conference on Learning Representations*, 2024. 2, 3
- [72] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemel 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. 16
- [73] Haoning Wu, Hanwei Zhu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Annan Wang, Wenxiu Sun, Qiong Yan, et al. Towards open-ended visual quality comparison. *arXiv preprint arXiv:2402.16641*, 2024. 3
- [74] 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. 5, 28, 29
- [75] 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. 16
- [76] 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. 28, 29
- [77] 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. 1, 5, 28, 29
- [78] 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. 15
- [79] 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. 3, 15
- [80] Yue Yang, Artemis Panagopoulou, Qing Lyu, Li Zhang, Mark Yatskar, and Chris Callison-Burch. Visual goal-step inference using wikihow. *arXiv preprint arXiv:2104.05845*, 2021. 1
- [81] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of LMMs: Preliminary explorations with GPT-4V (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1, 2023. 1
- [82] Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. MMT-Bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*, 2024. 1
- [83] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. MMMU: A massive multi-discipline multi-modal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 1
- [84] Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Haodong Duan, Songyang Zhang, Shuangrui Ding, et al. InternLM-XComposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023. 1
- [85] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, Songyang Zhang, Wenwei Zhang, Yining Li, Yang Gao, Peng Sun, Xinyue Zhang, Wei Li, Jingwen Li, Wenhui Wang, Hang Yan, Conghui He, Xingcheng Zhang, Kai Chen, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. InternLM-XComposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024. 5
- [86] Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. GPT-4V (ision) as a generalist evaluator for vision-language tasks. *arXiv preprint arXiv:2311.01361*, 2023. 3
- [87] 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. 16
- [88] Kaizhi Zheng, Xuehai He, and Xin Eric Wang. MiniGPT-5: Interleaved vision-and-language generation via generative vokens. *arXiv preprint arXiv:2310.02239*, 2023. 3, 5, 28, 29
- [89] 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. 3, 5, 25
- [90] 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. 15
- [91] 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. 16
- [92] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024. 1
- [93] 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. 17

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

- Sec. A includes the details of OpenING and data curation.
- Sec. B presents the details of evaluation and IntJudge.
- Sec. C provides the details of our experiments.
- Sec. D introduces the extended experimental results of training on the Dev Set of OpenING.
- Sec. E discusses the limitations of this study.

A. Details of OpenING and Data Curation

A.1. Hierarchical Structure of OpenING

We present in Table 6 all 56 specific tasks from OpenING, belonging to 23 meta-topics, including the number of Instances (# of Ins.), Meta-topic names, and capabilities of MLLMs evaluated in testing. Specifically, the 56 tasks are divided into 38 common tasks and 18 hard tasks. 38 common tasks are annotated by 28 professional annotators, instructed and monitored by 14 data experts. 18 hard tasks are annotated by 14 data experts as these tasks need certain domain knowledge for data extraction, collection, and annotation. It is noted that 18 hard tasks are extremely difficult, especially in collecting data from rare sources and annotating data into the desired format. The number of instances in each category is also provided.

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 corresponding text for each 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. Task Prompt Breakdown

We present a comprehensive breakdown of the task prompts used in our experiments. The tasks are designed to evaluate various capabilities of interleaved image-text generation methods, ranging from storytelling and creative design to problem-solving and interactive experiences. For each task, we provide the general prompt format and specific prompt examples to illustrate how the models are expected to generate interleaved image-text outputs in Table 10. 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, showing the generalization performance of methods more comprehensively.

A.4. 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.5. Task Abbreviations

Given the large number of tasks and methods tested in our benchmark, we use abbreviations for simplification in the manuscript. In order to facilitate understanding of possible abbreviations in the text, we provide respectively the abbreviations of the 23 meta-topics in Table 8 and the abbreviations of the 56 tasks in Table 9.

A.6. Interface of IntLabel

To ensure that annotations were consistent within the standard format across annotators, and to facilitate the annotation process, we developed a labeling tool for interleaved image-text data annotation, named IntLabel. All queries and answers in OpenING are annotated manually, checked and refined via our developed tool, IntLabel. The final saving format of the annotation is JSONL, alongside a folder containing the corresponding image. We present the GUI of our IntLabel in Fig. 10. IntLabel is developed based on PyQt⁵. We will also release the source code of IntLabel for contributing to the community.

A.7. Illustration of More Data Examples

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 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.

⁵<https://www.riverbankcomputing.com/software/pyqt>

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
Storyboard 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 Sources: ID number and data sources of each task.

Task ID	Task Name	Task Source
1.1	Story Writing	SEED-Story [79], StoryGen [40] and Storybird (https://storybird.com/read-picture-book)
1.2	Fiction Writing	SFGram [55]
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/R100004029).
2.3	Dynamic Biography Generation	Wikipedia (https://en.wikipedia.org/), Baidu Baike (https://baike.baidu.com/)
3.1	Storybook Completion	SEED-Story dataset [79], 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 [90] and DocLayNet [50]
4.2	Slide with Note Generation	ReIP [56] and manually collected in-house slides
5.1	Web GUI Navigation	GUI Odyssey [45] and GUI World [15]
5.2	In-APP GUI Navigation	GUI Odyssey [45] and GUI World [15]
5.3	Cross-APP GUI Navigation	GUI Odyssey [45] and GUI World [15]
5.4	OS GUI Navigation	GUI Odyssey [45] and GUI World [15]
6.1	Interactive Portrait Image Editing	InstructPix2Pix [10], PnPInversion [33] and SEED-Data-Edit [23]
6.2	Interactive Landscape Image Editing	InstructPix2Pix [10], PnPInversion [33] and SEED-Data-Edit [23]
7.1	Interior Design	Xiaohongshu (https://www.xiaohongshu.com/)
7.2	Architectural Design	Architecture Style Dataset [78] 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 [26], Package Design Dataset (https://www.kaggle.com/datasets/dagloxxkankwanda/package-design-dataset) and Xiaohongshu (https://www.xiaohongshu.com/)
7.5	Interactive Graphic Advertisement Editing	CreativeRanking [69] 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/p1/2024-07-11/doc-incidentfn8913329.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?xsec_token=AB5RmgUizZbLQLmLj8zWSmutvLdUpKq6gA30qz647fKv0&xsec_source=pc_search) and AI-generated content
10.5	Dream Analysis and Reconstruction	Zhougong's Dream Interpretation (https://m.zgjm.org/), in-house dream records and AI-generated content
11.1	Multimodal Biological Reasoning	CUB-200 [67] and Oxford 102 Flower [47]
11.2	Multimodal Landscape Reasoning	ADE20K [91] and Oxford 5k [3].
11.3	Multimodal Analogy Reasoning	MIRB [87] 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 [7] and OmniObject3D [75]
13.1	Interactive Multi-concept Image Composition	TVSum [57] 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 [12], AVA-Actions [27] and EPIC Kitchens [18]
16.2	Visual Traffic Forecasting	Argoverse [72], 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 [13]
18.2	Chat with Memes	MOD [21] 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 In-door Navigation	CARLA [19], VisDrone [93], Gibson Environment (http://gibsonenv.stanford.edu/) and Reverie (https://reverie.herokuapp.com/arXiv_Demo/)

Table 8. Abbreviation of Meta-topics.

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

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

Abbrev.	Task Name	Abbreviation	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 Mernes	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

Table 10. 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>
Storyboard 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 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>
Interactive Multi-concept Image Composition	<BEGIN> ***. Please summarize all the content in one image and write a blog post. <image> <image>	<BEGIN> This is a collection of four Christmas smoothies. Please summarize all the content in one image and write a blog post. <image> <image> <image> <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> <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 perceiving 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>

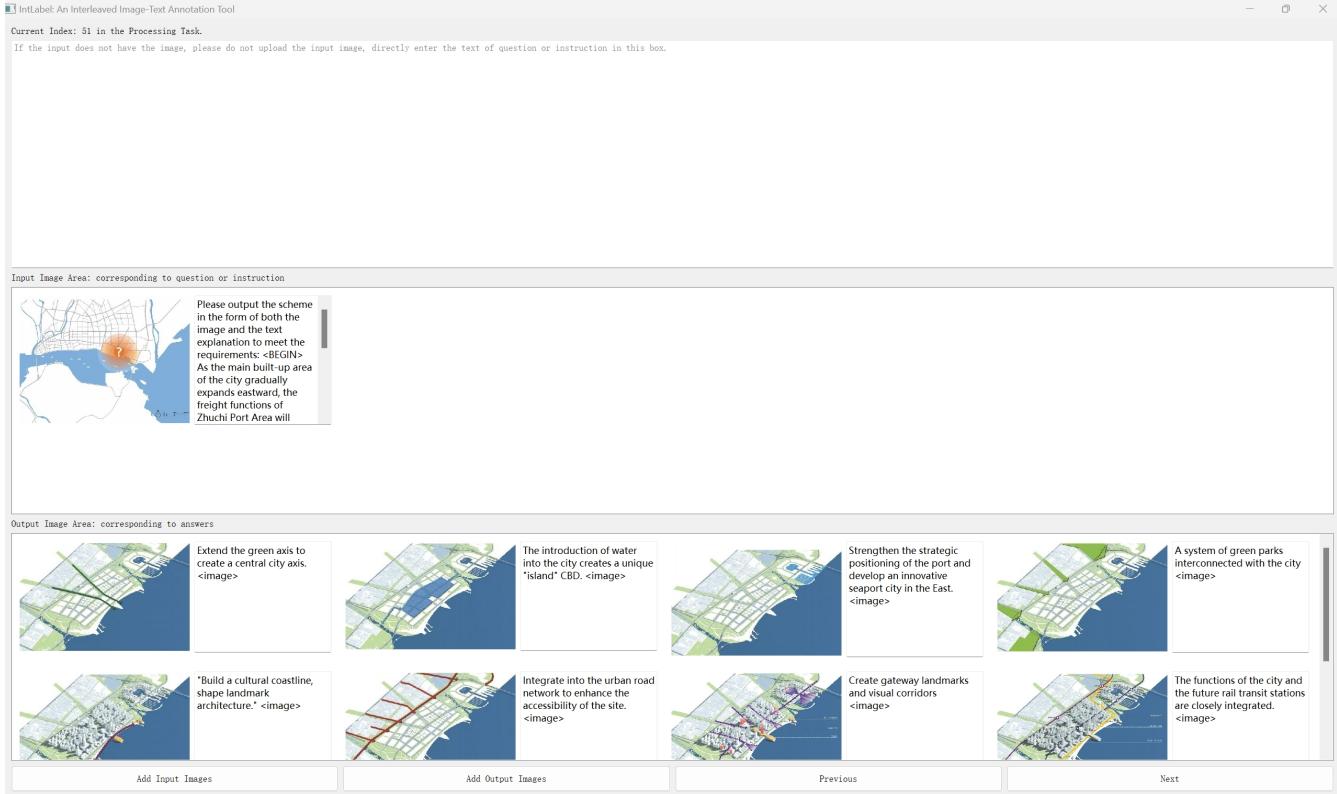


Figure 10. Interface of IntLabel, which shows a case where data is entered to finish annotation for an instance.

B. Details of Evaluation and IntJudge

B.1. 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, avoid-

ing content that may be offensive, inappropriate, or misleading.

- 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.

B.2. Prompt for GPT-based Judge

Fig. 11 presents the prompt used for a GPT-based judge tasked with comparing the quality of answers generated by two interleaved generation methods, named Model A and Model B, based on a given input question. The evaluation follows 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 and determine which model performed better, outputting a clear verdict such as "A is better," "B is better," or indicating a Tie (and choosing a more favorable method). This structured approach allows for a thorough, criterion-driven comparison of the two generated outputs, contributing to

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

a detailed understanding of the relative strengths of each model. The initial explorations on judges based on Qwen2-VL and InternLM-XComposer2.5 also adopt the same system prompt. In order to reduce the number of input tokens and save GPU memory when implementing these open-source MLLMs, we further refined these system prompts.

B.3. Prompt for Qwen, Intern and IntJudge

Fig. 12 illustrates the system prompts designed for obtaining optimal judgments based on Qwen2-VL, InternLM-XComposer2.5 and our IntJudge. This prompt was refined through extensive prompt engineering to maximize efficiency and reduce token usage, ultimately saving GPU memory when implementing these 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 well with human evaluators. We also provide a previously used prompt for Qwen2-VL and InternLM-XComposer2.5 for comparison in Fig. 12. The refined prompt allows for more streamlined input, ensuring the judgments are concise while still covering all essential evaluation aspects.

B.4. Prompt for GPT-based Scoring

Fig. 13 presents the system prompt designed for obtaining detailed scores from GPT-based evaluators. The prompt instructs GPT to evaluate interleaved image-text content based on seven key criteria. Each of these criteria is scored on a scale from 0 to 10, accompanied by a brief explanation of the assessment. The evaluation of GPT-based scoring aims to provide a supplementary analysis of the generated content, supporting further performance comparisons between models.

B.5. Annotation Interface of Interleaved Arena

The Interleaved Arena is introduced as an evaluation framework specifically designed to address the challenges of assessing open-ended interleaved image-text generation. Evaluating interleaved image-text generation is difficult due to: 1) There is a need to assess multiple images and text together; 2) There is no single correct answer (multiple solutions exist for an input). Since comparative evaluation has been shown to be more stable and reliable than subjective scoring [14], pairwise comparison is used 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 fairly select pairwise data from all available interleaved generation methods; and 2) An annotation interface for human judges to conduct evaluations manually. The annotation interface, shown in Fig. 14, is developed using PyQt and will be made available to researchers for collecting more manual

evaluation results, including through crowdsourcing. 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 [14, 89] 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.6. Training Losses of IntJudge

The training of IntJudge utilized a comprehensive loss framework to enhance its evaluation capabilities. As defined in the main text, the total loss combined four components with weights coefficients: cross-entropy loss \mathcal{L}_{CE} for language modeling and alignment, contrastive loss \mathcal{L}_{CT} to align image and text embeddings, MSE loss \mathcal{L}_{MSE} for reducing prediction errors, and pairwise ranking loss \mathcal{L}_{PR} to prioritize correct rankings of outputs, detailed as follows.

Cross-entropy loss for language modeling is given by:

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

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

Contrastive Loss for aligning image and text embeddings is written as:

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

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.

Mean Squared Error Loss for image feature regression is given by:

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

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 is given by:

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

where $f(x_i^+)$ and $f(x_i^-)$ represent the scores assigned to positive and negative examples, respectively. The combined

Final 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)":

Former 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 12. The system prompts for using MLLMs as a judge to compare outputs of two interleaved generation methods.

You are an expert in multimodal content evaluation. Your task is to assess the interleaved image-text content generated by a model. The model's input, starting with "INPUT:", can be a question or an instruction requesting the generation of the interleaved content, which may include images to offer visual information. Its output, starting with "OUTPUT:", can be the generated answer that includes interleaved images and texts based on the given input.

****Image Representation:****

- The image of 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).
- In this evaluation, you can accept image inputs directly.

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

****Evaluation Criteria:****

1. **Correctness**: Evaluate the accuracy and validity of the generated content corresponding to the given input. This includes assessing the logical coherence, factual accuracy, and presence of the text, as well as considering the reasonableness, fidelity, and presence 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 importance 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.

****Scoring Guidelines:****

- **0-2**: Major deficiencies
- **3-5**: Minor issues
- **6-8**: Good performance with minor improvements needed
- **9-10**: Exemplary performance

****Instructions:****

- For each criterion, provide a brief assessment and assign a score on a scale from 0 to 10, following the Scoring Guidelines.
- Structure your response in the following JSON format:

```
{  
  "scores": {  
    "Correctness": {  
      "Score": 0-10,  
      "Justification": "Brief explanation of any issues identified"  
    },  
    "Image-Text Coherency": {  
      "Score": 0-10,  
      "Justification": "Brief explanation of any issues identified"  
    },  
    "Multi-step Consistency": {  
      "Score": 0-10,  
      "Justification": "Brief explanation of any issues identified"  
    },  
    "Content Quality": {  
      "Score": 0-10,  
      "Justification": "Brief explanation of any issues identified"  
    },  
    "Human Preference Alignment": {  
      "Score": 0-10,  
      "Justification": "Brief explanation of any issues identified"  
    },  
    "Completeness": {  
      "Score": 0-10,  
      "Justification": "Brief explanation of any issues identified"  
    },  
    "Content Richness": {  
      "Score": 0-10,  
      "Justification": "Brief explanation of any issues identified"  
    }  
  }  
}
```

- Be objective and thorough in your evaluation, providing clear justifications for your scores.
- Remember that you can accept image inputs directly, so you should analyze the images for each criteria.

****Data to Review:****

Figure 13. The system prompt for obtaining detailed scores from GPT-based evaluators. Brief explanations are also required to support further performance analysis of different models.

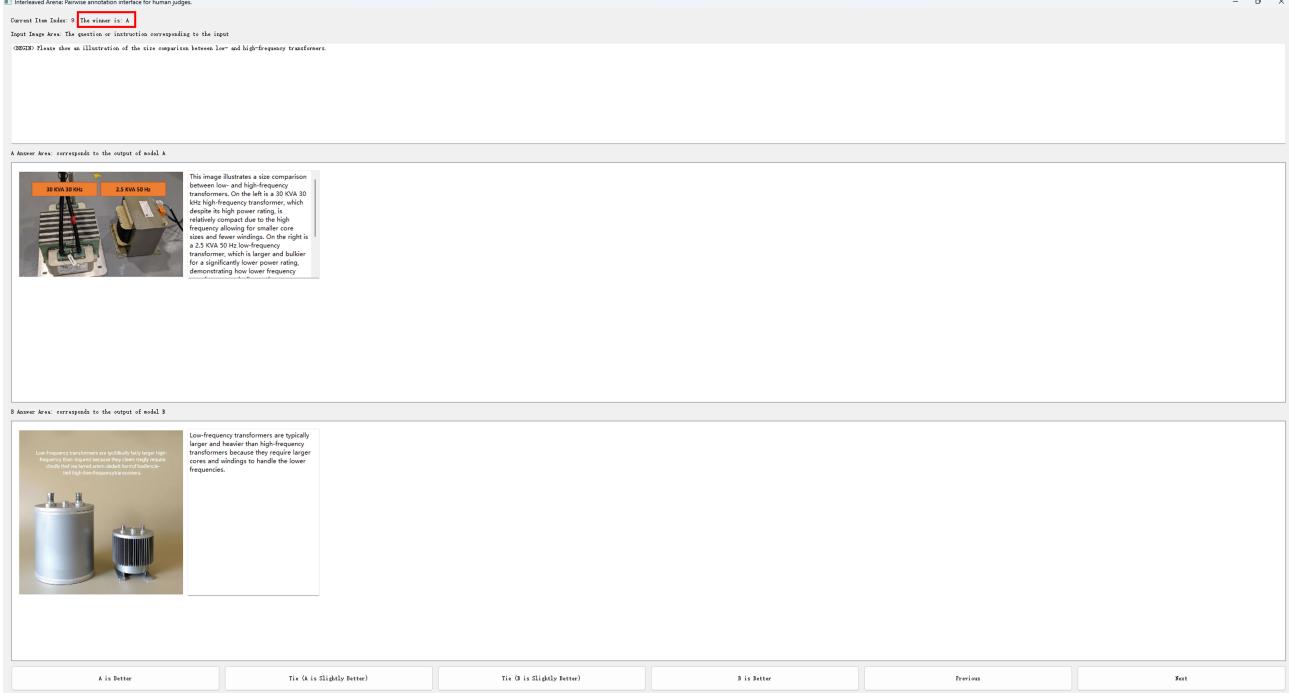


Figure 14. Annotation Interface of Interleaved Arena for human judges to compare the anonymous outputs of model A and model B. Human evaluators are instructed to select "A is Better" or "B is Better" to choose a winner for the pairwise comparison. When the two outputs are similar in quality and a tie 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.

total loss ensures language generation basis, multimodal understanding capabilities, ranking accuracy and consistency across similar inputs during training.

C. Details of Experiments

C.1. Baseline Details

We evaluate 12 representative methods, grouping them into three categories: **1) Integrated pipeline**, which involves separate models for text and image generation in two stages, such as GPT-4o+DALL-E3 (DALL-E3) [8, 48] and Gemini1.5+Flux [9, 63]; **2) Two-stage generator**, which employs a unified model architecture to produce text and images in separate stages, including Emu3 [70], VILA-U [76], Emu2 [59], SEED-X [23], and Show-o [77]; and **3) End-to-end generator**, which directly generates image-text outputs in a single step, such as GILL [35], NExT-GPT [74], MiniGPT-5 [88], SEED-LLaMA [22], and Anole [16]. 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.

For the Integrated Generation Pipeline, the advantage lies in its modularity, allowing each component to specialize in

its respective task—text generation or image creation. This approach can leverage the strengths of SOTA commercial 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, Gemini 1.5 combined 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, as with other pipeline methods, synchronization and contextual consistency between stages remain areas for further improvement.

In the 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.

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 [62]. However, the open-sourced version of 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 tackling the current challenges of interleaved multimodal generation, ranging from modular pipelines to unified architectures. This taxonomy allows for a comprehensive evaluation of different approaches and analyzes potential pathways for developing MLLMs in this domain. We detail below the description of each model in the respective categories:

- **Integrated Pipelines:**

- **GPT-4o+DALL-E·3** [8, 48]: This pipeline leverages GPT-4o [48] to generate text and captions for the desired image generation. The captions are subsequently fed into DALL-E·3[8] to produce the corresponding images. The final output combines the text and images in their original sequence, enabling multimodal content generation through a staged process.
- **Gemini1.5+Flux** [9, 63]: This method integrates Gemini1.5 Pro, a powerful LMM, for text generation and Flux-schnell, a fast and efficient image generation model. The pipeline emphasizes high-quality and coherent text-to-image alignment through a structured two-step process.

- **Two-stage Generators:**

- **Emu2** [59]: Emu2 is a 37B MLLM with multimodal generation capabilities. The pretrained Emu2 is finetuned 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** [23]: 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** [77]: Show-o is a unified Transformer model combining autoregressive and diffusion approaches to flexibly handle multimodal understanding and gener-

ation tasks. We implement Show-o adopts a similar two-stage generation approach, focusing on separately producing interleaved multimodal content step by step.

- **Emu3** [70]: 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** [76]: 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** [88]: 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 as MiniGPT-5MMD.
- **GILL** [35]: 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** [74]: 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** [22]: 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** [16]: Anole is an end-to-end interleaved generation model fine-tuned on Chameleon, leveraging pre-

trained 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.

C.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.

C.3. More Experimental Results on New Models

We present more experimental results in Table 11. 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 11 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.

C.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

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 11. 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.

include more domain knowledge may mitigate these issues and improve their interleaved generation performance.

C.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.

C.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.

C.7. Case Study

We present a case study analysis of the interleaved content generated by 14 methods on various tasks on OpenING in Fig. 19, which illustrates 23 pairs of model battle results across 23 meta-topics. The gold medal represents the winner

of a pairwise comparison and the silver medal denotes a more favorable output in a tie scene. The meta-topic name and task name for each pairwise data instance are given, which are separated by colons. We classify the errors in the generated content into the following eight categories. The breakdown of these eight error types is as follows:

1. **No Text or Image:** The model fails to generate either text or images when they are expected. This includes situations where no text is produced or no image is provided when required.
2. **Factual Error:** The model provides incorrect information, repeats answers, fails to follow instructions or refuses to answer, misunderstands the input, or demonstrates reasoning errors, including generating wrong images.
3. **Content Incoherent:** The generated content is incoherent in style or entities, either with the input or between multiple outputs. This involves cases where the style or entities do not match the input or are inconsistent across outputs.
4. **Offensive Content:** The output contains uncomfortable material or raises safety concerns due to inappropriate content, such as violent, harmful, or otherwise unsafe elements that violate ethical guidelines.
5. **Image-Text Inconsistent:** The images and accompanying text do not correspond or align appropriately, leading to confusion or misinterpretation.
6. **Poor Image Quality:** The generated images are of low quality, such as being completely black or blank, blurred, or unrealistically rendered.
7. **Poor Text Quality:** The text is of low quality, including nonsensical mumbling, grammatical issues that hinder readability and understanding, or being too short to convey meaningful information.

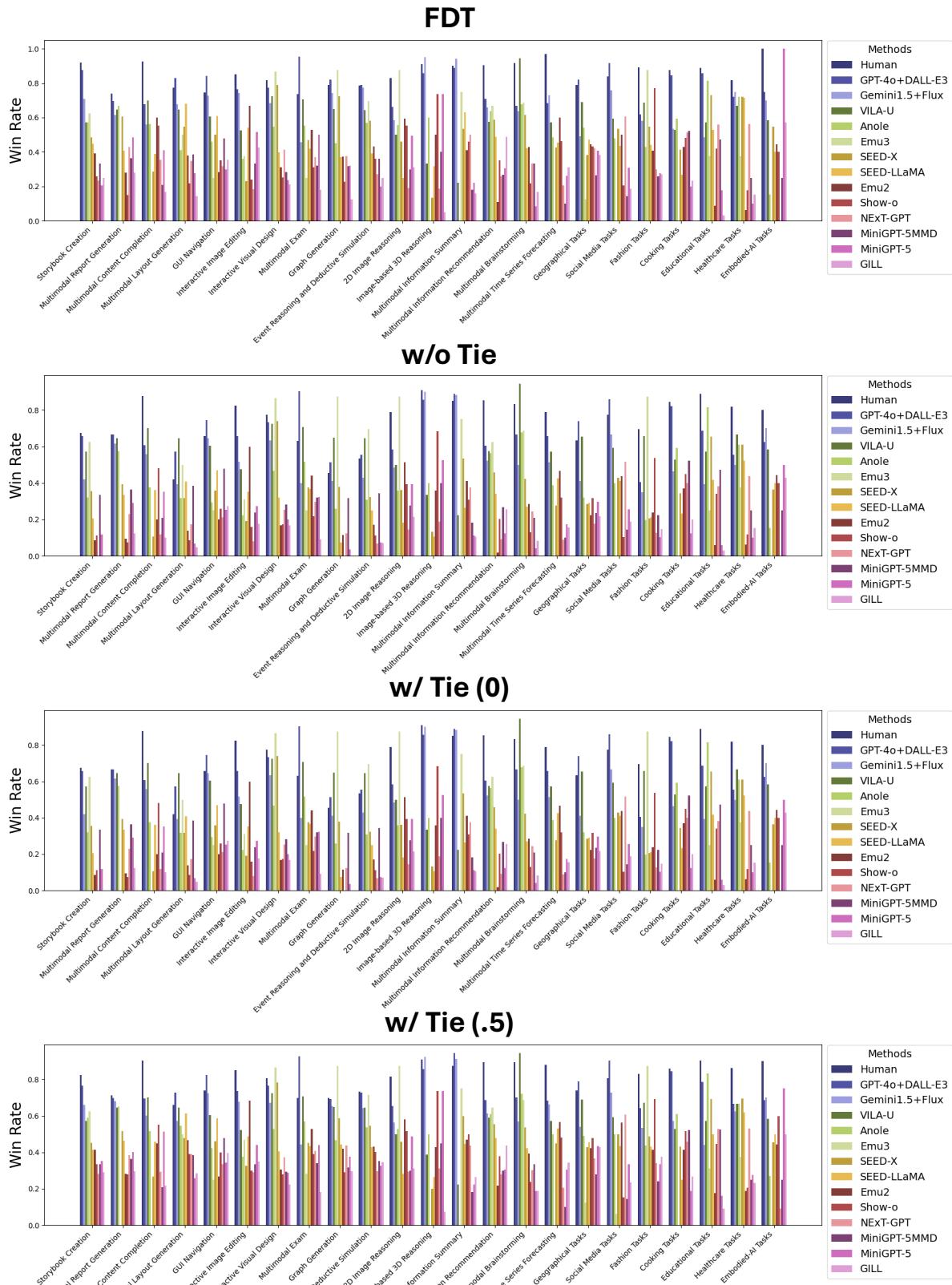


Figure 15. The win rates of 14 interleaved generation methods across 23 meta-topics.

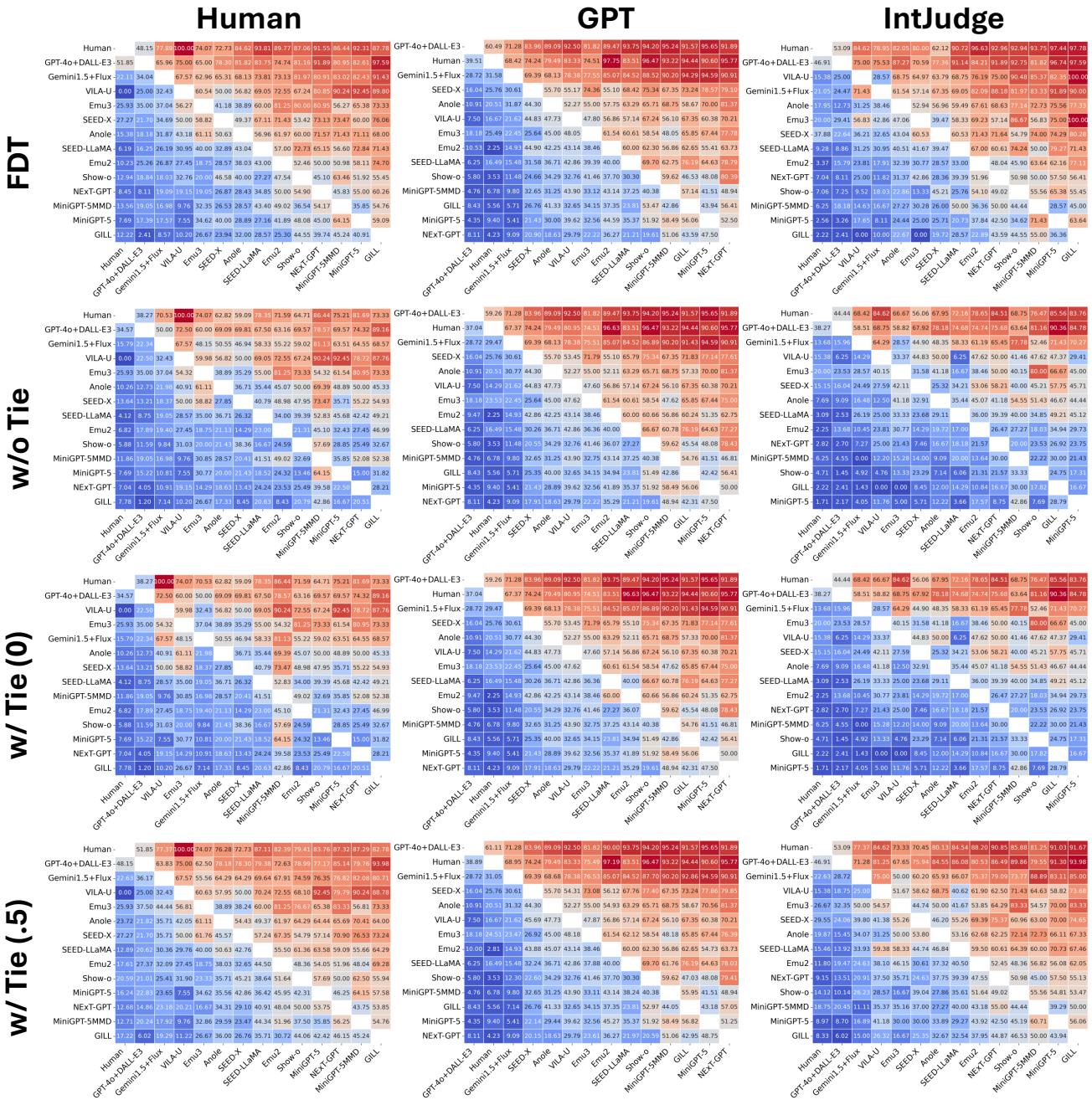


Figure 16. Win rate matrices of 14 interleaved generation methods, evaluated by Human, GPT-4o, and our IntJudge, respectively.

8. **Incomplete Response:** The textual output stops abruptly, or outputs an incomplete or truncated response.

We introduce 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

D. Extension of OpenING

verify if finetuning on the specific data of OpenING can improve 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 com-

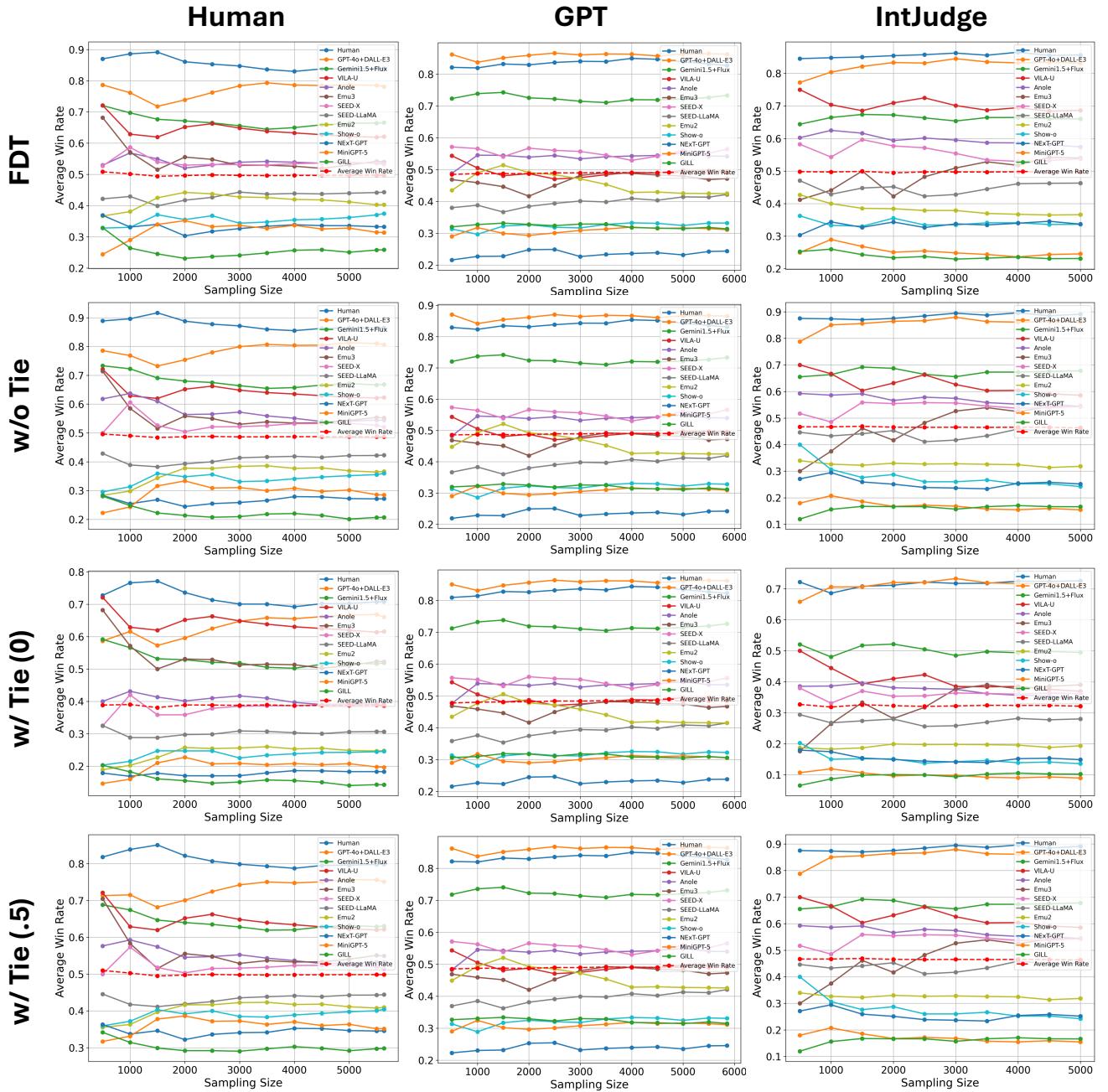


Figure 17. Win rate curves with respect to different sampling sizes.

putations to speed up the training process. The results are evaluated on the Test Set of OpenING using IntJudge.

In Table 12, 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 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.

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 12. 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.

These findings confirm that training on a specialized interleaved image-text dataset such as the Dev Set of OpenING enhance 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.

E. Limitations of This Study

Although the OpenING benchmark represents a significant step forward in evaluating interleaved image-text generation, it faces several limitations that present opportunities for improvement. First, while OpenING expands task diversity with 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, though the IntJudge model improves alignment with human evaluations, its generalizability is still constrained 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 both labor-intensive and costly. While the proposed 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. What is more, 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.

Figure 18. Representative examples from 23 Meta-Topics in Our OpenING Benchmark.

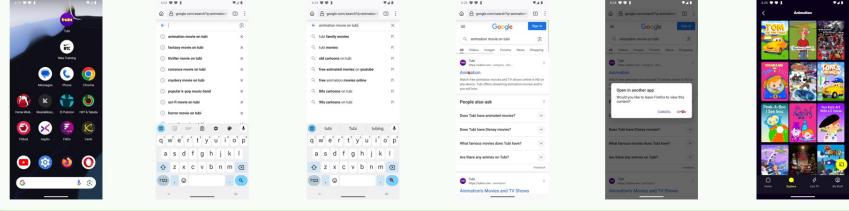
GUI Navigation

Please give the results of GUI navigation with interleaved 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 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

<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.



Interactive 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>



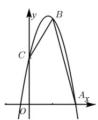
The modified image shows ten individuals outdoors on a grassy landscape, with the surroundings remaining unchanged. The most noticeable alteration is the person on the far left, now dressed in an orange and white plaid shirt, wearing sunglasses and a light blue cap, and holding two yellow cocktails.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



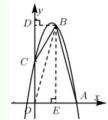
Multimodal Exam



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, $\angle A = \angle AHB$. Thus, $\triangle DAE \cong \triangle HBE$ (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>



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>

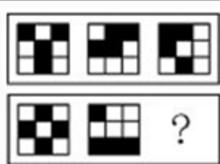
The diet of mosquitoes <image> The three most common types of mosquitoes that bite humans in China. <image> The steps of a mosquito sucking blood <image> The flying method of mosquitoes <image> The Dangers of Mosquitoes <image> The Differences and Similarities Between Mosquitoes and Flies <image> Knowledge of mosquitoes <image>



2D 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 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



<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



<BEGIN> This is a house. Its appearance features are:\n"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. [\(image\)](#)



Multimodal Information Summary



<BEGIN> My Recent Week of Brunch, Please help me summarize the Brunches I've made in a single image and accompany text for this blog [\(image\)](#) [\(image\)](#) [\(image\)](#) [\(image\)](#) [\(image\)](#) [\(image\)](#)



The Healing Magic of Life\nTake Good Care of Your Energetic Self\nA Week's Light Breakfast and Lunch Records\nCocktail Shrimp\nSoaking Vegetables in Broth\nDaily Energy Bowl\nCitrus Salad\nOne-Pot Air Fryer Meal\nFree Combination Brunch [\(image\)](#)



Multimodal Information 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\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. [\(image\)](#)

Coco\nSynopsis: 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

 <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

 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>](#) [<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

 <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>](#) [<image>](#)



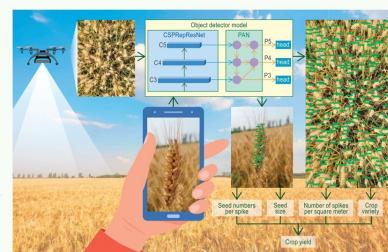
 Certainly, here is the English translation:
"Alright, based on the image you provided, we have generated a render of how the clothes will look on you. The result after changing outfits is not satisfactory. The pattern on the chest of the white top is too vivid, which conflicts with the style of the pants.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



Multimodal Brainstorming

 <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:
Data Acquisition: Drones 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.
Smartphones: Used by individuals on the ground to take close-up photos of the crops for more detailed analysis.
Data Processing:
Input Images: The images collected by both drones and smartphones are fed into an object detection model.
Object 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.
Feature Extraction: Extract specific features from the crops, such as the number of spikes per square meter and seed size, which are crucial for yield estimation.
Data Analysis:
Quantitative Analysis: Using the detected features (e.g., seed numbers per spike, number of spikes per square meter, seed size) to estimate the crop yield.
Crop Variety Identification: Identify the variety of the crop based on physical features detected in the imagery.
Output:
Crop Yield Estimation: Calculate estimated yield based on the analyzed data.
Visualization: Provide a user-friendly interface (possibly on the smartphone app) to display the analyzed data and yield predictions.
Feedback Loop:
Adjustments and Calibration: Adjust the model based on ground-truth data collected manually or feedback from the users to improve accuracy over time.
This framework would be beneficial for agricultural scientists, farmers, and agritech companies to monitor crop health and predict yields nondestructively and efficiently. [<image>](#)



Cooking Tasks



<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>



Social Media Tasks



<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 freedom. <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



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>



Embodied AI Tasks



<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> <image> <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>

Healthcare Tasks



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>



Figure 19. Illustration of 23 pairs of generated outputs across 23 meta-topics. The gold medal represents the winner of pairwise comparison and the silver medal denotes a more favorable output in a tie scene of the pairwise comparison. The meta-topic name and selected task name for each pairwise data instance are given, which are separated by colons.

Story Writing: Storybook Creation



<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

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 despaching-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>



Museum Guide Book Generation: Multimodal Report Generation



<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>



Content Incoherent

Poor Image

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>



Content Incoherent

Poor Text

Incomplet

Multimodal Report Completion: Multimodal Content 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>

X Content Incoherent

Poor Image

Model B (Gemini1.5+Flux)

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



X Content Incoherent

Poor Text

Incomplete

Slide with Note Generation: Multimodal Layout 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>



X Factual Error

Poor Image

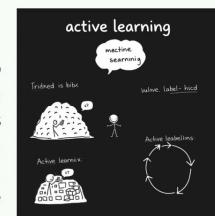
Poor Text

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>



Cross-APP GUI Navigation: 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

Poor Image

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>



Content Incoherent

Image-Text Inconsistent

Interactive Landscape Image Editing: Interactive 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

Content Incoherent

Model B (GILL)

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



No Text or Image

Factual Error

Architectural Design: Interactive Visual Design



<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

Model B (SEED-X)

1. Classical-style scenic area design ideas:

- Use classical architecture elements such as columns, arches, and pediments to create a sense of grandeur and elegance.
- Incorporate natural elements such as fountains, statues, and gardens to create a peaceful and relaxing atmosphere.
- 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:

- Classical-style building with columns, arches, and pediments. <image>



Circuit Problem Test: Multimodal Exam



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.



Poor Image

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.



Factual Error

No Text or Image

Figure Relationship Diagram Generation: Graph 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

1

| --- | --- |
| 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

Model B (Gemini1.5+Flux)

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>



Dream Analysis and Scene Reconstruction: Event Reasoning and Deductive Simulation

 <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)



Factual Error

 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>



Interactive Jigsaw Puzzle: 2D Image Reasoning



<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 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>



Interactive Novel View Synthesis: Image-based 3D Reasoning

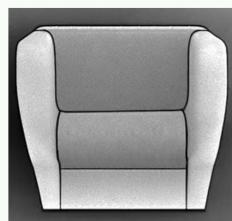


<BEGIN> This is a sofa. Its appearance features are: "This is a single-seat sofa, the whole is off-white, a little yellowish, the sofa is made of many folds, these folds are designed rather than due to prolonged use, the overall structure is axisymmetric". Can you guess what it looks like from the side?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



Model A (MiniGPT-5MMD)

I will draw the side view. <image>



Factual Error

Content Incoherent

Poor Image

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>



Interactive Multi-concept Image Composition: Multimodal Information Summary



<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><image><image><image>



Model A (GILL)

tea cup and a book. <image>



Content Incoherent

Poor Text



Model B (Emu2)

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



Poor Text

Poor Image

Interactive Film and Television Recommendation: Multimodal Information 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?



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)

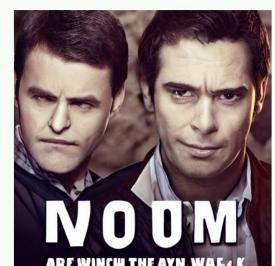
Dive into a world where Brazilian folklore meets urban fantasy in this visually stunning series. <image>



Poor Image

Model B (GILL)

I would recommend the following:
<image>



Poor Text

Incomplete

Academic Scenarios Brainstorming: Multimodal Brainstorming



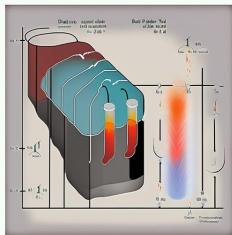
<BEGIN> How do temperature and concentration gradients evolve over time in a star-finned adsorbent bed used in adsorption chillers? Please show an image for illustration.



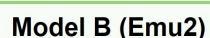
Model A (SEED-X)

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>

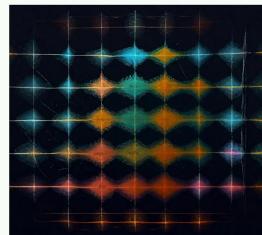


Incomplete



Model B (Emu2)

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



Factual Error

Multimodal Action Anticipation: Multimodal Time Series Forecasting



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>



Content Incoherent

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>



Interactive Street View Image Rendering: Geographical Tasks



<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 B (GILL)

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



Factual Error

Poor Image

Chat with Memes: Social Media Tasks



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>



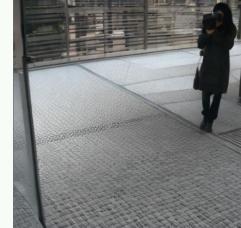
Model A (GPT-4o+DALL-E3)

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



Model B (MiniGPT-5)

Where are we going ? <image>



Factual Error

Fashion Trend Forecasting: Fashion Tasks



<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 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

Poor Image

Multimodal Cooking Assistant: Cooking Tasks



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>



Poor Text

Poor Image

Model B (Gemini1.5+Flux)

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



Image-Text Inconsistent

Interactive Science Popularization: Educational Tasks

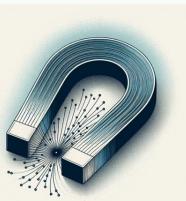


<BEGIN> What is the magnetic field of a horseshoe magnet like? Please explain using diagrams and illustrations. <image>



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>

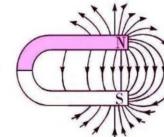


Factual Error



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>



Fitness and Health Consulting: Healthcare Tasks



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

Poor Text

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>



Content Incoherent

Autonomous Driving and In-door Navigation: Embodied-AI Tasks



<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><image><image>



Model A (GILL)

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



Content Incoherent

Model B (MiniGPT-5MMD)

Here is the output.



No Text or Image