Holistic Evaluation for Interleaved Text-and-Image Generation

Anonymous ACL submission

Abstract

Interleaved text-and-image generation has been an intriguing research direction, where the models are required to generate both images and text pieces in an arbitrary order. Despite the emerging advancements in interleaved generation, the progress in its evaluation still significantly lags behind. Existing evaluation benchmarks do not support arbitrarily interleaved images and text for both inputs and outputs, and they only cover a limited number of domains and use cases. Also, current works predominantly use similarity-based metrics which fall short in assessing the quality in open-ended scenarios. To this end, we introduce INTERLEAVEDBENCH, the first benchmark carefully curated for the evaluation of interleaved text-and-image generation. INTERLEAVEDBENCH features a rich array of tasks to cover diverse real-world use cases. In addition, we present INTERLEAVEDEVAL, a strong reference-free metric powered by GPT-4o to deliver accurate and explainable evaluation. We carefully define five essential evaluation aspects for INTERLEAVEDEVAL, including *text quality*, *perceptual quality*, *image coherence*, *text-image coherence*, and *helpfulness*, to ensure a comprehensive and fine-grained assessment. Through extensive experiments and rigorous human evaluation, we show that our benchmark and metric can effectively evaluate the existing models with a strong correlation with human judgments surpassing previous reference-based metrics. We also provide substantial findings and insights to foster future research in interleaved generation and its evaluation.[1]

1 Introduction

Multimodal learning has been a rapidly developing research field given the recent advancements in Large Multimodal Models (LMMs) (Xu et al., 2023; Dai et al., 2023; Liu et al., 2023). While these models can perform diverse tasks such as detailed image description and visual question answering, the outputs are limited to the text-only format, which hinders their broader applications. More recently, there has been a growing focus on enhancing LMMs with the capability of *interleaved generation*, i.e., generating multimodal content that seamlessly integrates both text and one or multiple images (Koh et al., 2023; Dong et al., 2024; Sun et al., 2023b,a). This opens new avenues for applications in diverse challenging scenarios, such as creative content generation (Anantrasirichai and Bull, 2022), visual storytelling (Huang et al., 2016; Lukin et al., 2018), and multimodal script generation (Yang et al., 2021; Qi et al., 2024).

While the LMMs for interleaved generation are continuously gaining stronger capabilities, progress in the *evaluation* of interleaved generation significantly lags behind with several critical challenges remaining. **First**, most existing works for interleaved generation quantitatively benchmark the models on text-to-image tasks where the output is usually one single image (Koh et al., 2023; Dong et al., 2024). However, such evaluation methods would fail to assess model performance in the real-world scenarios of interleaved generation, where the output usually consists of interleaved text and images. **Second**, apart from human evaluation which is costly and time-consuming, existing works still heavily rely on reference-based metrics such as BLEU (Papineni et al., 2002) FID (Heusel et al., 2017) that measure the similarity between generated samples and gold references. Such similarity-based metrics often fail to accurately capture outputs' quality, especially in open-ended tasks such as creative generation and visual storytelling. **Third**, the evaluation of interleaved generation is complex and involves many different aspects, such as *perceptual quality*, *coherence* between text and images, and *helpfulness* of the overall content. One single aspect is usually insufficient to reflect the

[1] The source code and datasets will be publicly available for research purposes.

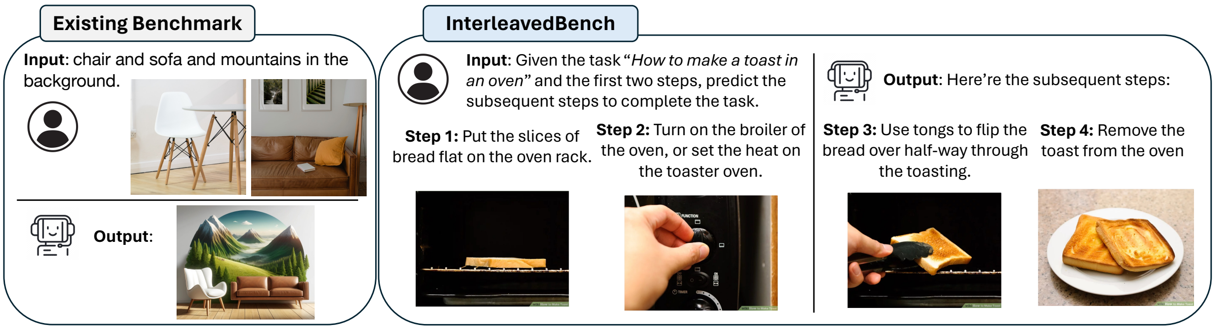

Existing Benchmark
Input: chair and sofa and mountains in the background.
Output:
InterleavedBench
Input: Given the task "How to make a toast in an oven" and the first two steps, predict the subsequent steps to complete the task.
Step 1: Put the slices of bread flat on the oven rack.
Step 2: Turn on the broiler of the oven, or set the heat on the toaster oven.
Output: Here're the subsequent steps:
Step 3: Use tongs to flip the bread over half-way through the toasting.
Step 4: Remove the toast from the oven

Figure 1: Comparison between the existing benchmark (multi-concept image composition (Kumari et al., 2023a)) and our INTERLEAVEDBENCH. Compared with the existing benchmark, INTERLEAVEDBENCH has the following features: (1) both input and output can have arbitrarily interleaved text and images, and (2) each instance has a detailed instruction to benchmark models' instruction-following capability.

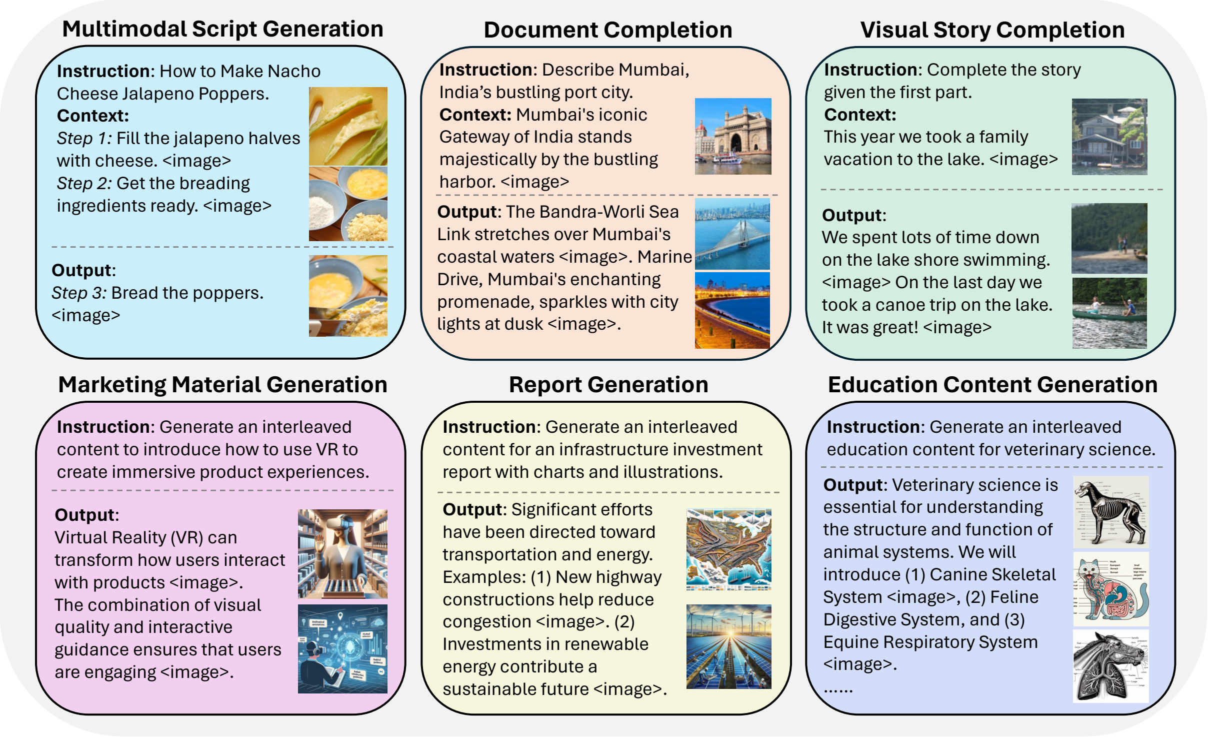
overall quality. For example, despite the images in one output having good perceptual quality, the output can still be not helpful to users if the generated content is not coherent with the context, e.g., the request from users.

To address these critical limitations, we introduce INTERLEAVEDBENCH, the first benchmark for holistic evaluation of interleaved text-and-image generation. We construct INTERLEAVEDBENCH with a high-quality and diverse collection of interleaved generation scenarios that encompass a wide range of real-world use cases, including creative generation, multimodal script generation, visual storytelling, and many others. We compare our INTERLEAVEDBENCH and one existing benchmark (Kumari et al., 2023b) closest to our dataset in Figure 1. To support the evaluation, we also introduce INTERLEAVEDEVAL, a strong reference-free evaluation metric based on GPT-4o (OpenAI, 2024), the current state-of-the-art LMM. INTERLEAVEDEVAL can take in any evaluation instructions and provide a fine-grained evaluation along with detailed explanations. We carefully curate a multi-aspect evaluation criterion to ensure a holistic evaluation for INTERLEAVEDEVAL. Specifically, we define five essential aspects for interleaved evaluation, including *text quality, perceptual quality, image coherence, text-image coherence*, and *helpfulness*, following the principles that (1) these aspects are generally applicable in different scenarios, (2) these aspects are atomic and orthogonal to each other, and (3) the combination of these aspects can comprehensively cover the critical dimensions in interleaved generation.

Extensive experiments and rigorous human evaluation demonstrate that **(1)** Our curated INTERLEAVEDBENCH posts unique and significant challenges to the existing integrated LMMs (e.g., GILL (Koh et al., 2023) and EMU-2 (Sun et al., 2023a)) for interleaved generation, where the quality of their outputs are far from satisfying. The pipeline systems combined with a strong LMM (e.g., GPT-4o) and a separate image generation model (e.g., DALLE3 (Betker et al.)) generally achieve better results but still struggle on certain tasks; **(2)** INTERLEAVEDEVAL can achieve a good correlation with human judgments with significant improvement over previous automatic evaluation metrics; **(3)** The evaluation of interleaved generation remains a very challenging direction due to its complexity and the limitation of the existing LMM-based evaluator. We believe that our work can provide useful resources and insights for interleaved generation and its evaluation.

2 Related Work

Large Multimodal Models for Interleaved Generation The advent of large multimodal models (LMMs) (Koh et al., 2023; Sun et al., 2023a) has significantly advanced the field of interleaved text-and-image generation. Previous models such as DALL-E (Ramesh et al., 2021) and Stable Diffusion (Podell et al., 2023) have demonstrated impressive capabilities in generating high-quality images conditioned on textual descriptions. However, previous focus has predominantly been on unidirectional generation tasks, either from text to image or image to text, without considering the interleaved generation scenarios where text and images are seamlessly integrated within the same output. Recent works have begun to address this gap, with the LMMs extended with diffusion models such as GILL (Koh et al., 2023), EMU (Sun et al., 2023b), and DreamLLM (Dong et al., 2024),



Multimodal Script Generation

Instruction: How to Make Nacho Cheese Jalapeno Poppers.
Context:
Step 1: Fill the jalapeno halves with cheese. <image>
Step 2: Get the breading ingredients ready. <image>

Output:
Step 3: Bread the poppers. <image>

Document Completion

Instruction: Describe Mumbai, India's bustling port city.
Context: Mumbai's iconic Gateway of India stands majestically by the bustling harbor. <image>

Output: The Bandra-Worli Sea Link stretches over Mumbai's coastal waters <image>. Marine Drive, Mumbai's enchanting promenade, sparkles with city lights at dusk <image>.

Visual Story Completion

Instruction: Complete the story given the first part.
Context:
This year we took a family vacation to the lake. <image>

Output:
We spent lots of time down on the lake shore swimming. <image> On the last day we took a canoe trip on the lake. It was great! <image>

Marketing Material Generation

Instruction: Generate an interleaved content to introduce how to use VR to create immersive product experiences.

Output:
Virtual Reality (VR) can transform how users interact with products <image>. The combination of visual quality and interactive guidance ensures that users are engaging <image>.

Report Generation

Instruction: Generate an interleaved content for an infrastructure investment report with charts and illustrations.

Output: Significant efforts have been directed toward transportation and energy. Examples: (1) New highway constructions help reduce congestion <image>. (2) Investments in renewable energy contribute a sustainable future <image>.

Education Content Generation

Instruction: Generate an interleaved education content for veterinary science.

Output: Veterinary science is essential for understanding the structure and function of animal systems. We will introduce (1) Canine Skeletal System <image>, (2) Feline Digestive System, and (3) Equine Respiratory System <image>.
......

Figure 2: Illustration of examples in our INTERLEAVEDBENCH from six representative use cases.

exploring the generation of mixed text and image outputs. These models leverage advanced architectures and training techniques to enhance their ability to produce coherent and contextually relevant interleaved content. Despite these advancements, the evaluation of such models remains an underexplored area, with most evaluations still relying on separate assessments of text and image quality or simplistic reference-based metrics. Our proposed INTERLEAVEDBENCH benchmark aims to bridge this gap by providing a holistic evaluation framework tailored specifically for interleaved text-and-image generation.

Evaluation of Multimodal Generation Tasks Evaluating multimodal generation tasks presents unique challenges due to the inherent complexity of assessing both textual and visual components simultaneously. Traditional metrics for text generation, such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005), fall short when applied to multimodal outputs as they fail to capture the visual quality and coherence with textual content. Similarly, visual generation metrics like FID (Heusel et al., 2017) and IS (Salimans et al., 2016) are inadequate for evaluating the textual elements accompanying the images. To address this, recent studies have employed multimodal metrics such as CLIPScore (Hessel et al., 2021), which leverages the alignment capabilities of the CLIP model to measure the similarity between generated images and their corresponding textual descriptions. However, CLIPScore can only measure the alignment between text and images, which is not sufficient to evaluate the quality of generated output comprehensively. Moreover, human evaluations, although more reliable, are resource-intensive and cannot be scalable. Our INTERLEAVEDBENCH benchmark introduces a novel approach to evaluate interleaved text-and-image generation by incorporating multiple aspects of quality assessment, thus providing a more nuanced and holistic evaluation framework.

3 INTERLEAVEDBENCH

We introduce INTERLEAVEDBENCH, the first comprehensive benchmark meticulously constructed to evaluate text-and-image interleaved generation. Figure 2 shows some examples from INTERLEAVEDBENCH.

3.1 Dataset Curation Process

Our dataset includes two subsets: a **context-based** subset where the instances contain a multimodal context of interleaved text and images in the input (first row in Figure 2), and a **context-free** subset with text-only inputs (second row in Figure 2). The context-free subset can assess whether

Dataset Name	Detailed Instruction	Image Input	Text Output	Image Output
MagicBrush (Zhang et al., 2023)	No	Single	No	Single
DreamBench (Chen et al., 2024)	No	Multiple	No	Single
CustomDiffusion (Kumari et al., 2023a)	No	Multiple	No	Single
DreamEditBench (Li et al., 2023)	No	Multiple	No	Single
Mantis-Eval (Jiang et al., 2024)	Yes	Multiple	Yes	No
InterleavedBench (Ours)	Yes	Multiple	Yes	Multiple

Table 1: Comparisons between InterleavedBench and existing open-sourced multimodal evaluation benchmarks. The highlighted features of our benchmark include detailed instructions and multiple images in input and/or output that are arbitrarily interleaved with text.

the model can creatively generate interleaved content based on the text-only instruction, while the context-based subset can better benchmark the coherence and consistency of generated outputs.

Collection of Context-based Subset **Firstly**, we collect the source data of the context-based subset from existing academic datasets or web resources. Specifically, we collect the data of multimodal script generation from WikiHow (Yang et al., 2021), visual story completion from VIST (Huang et al., 2016), activity generation from the dense captions and the extracted video frames in ActivityNet Captions (Krishna et al., 2017), sequential image editing from MagicBrush (Zhang et al., 2023), and multi-concept image composition from CustomDiffusion (Kumari et al., 2023a). For web resources, we apply an automatic data filtering pipeline to discard the samples with poor quality to obtain a small set of source data. We detail our data filtering pipeline in Appendix A. **Secondly**, after collecting the source data (either from academic benchmarks or web resources), we then apply a human selection process to manually select the samples based on data quality and diversity (i.e., avoiding selecting similar samples). **Finally**, we ask human experts to annotate an instruction I for each sample based on the collected content. We include the details of the data selection and instruction annotation process in Appendix A. For the samples that are originally interleaved articles, we pick the first k images and their associated text as the *context* $\mathcal{C}$ for the input. k is randomly sampled for each example and ranges from 1 to the maximum number of images minus 1 since we need to ensure the output contains at least one image. The rest of the images and text are used as the gold reference.

Collection of Context-free Subset The context-free subset consists of the use cases of *marketing material generation, report generation, education content generation*, and *fairytale generation* as they are common and practical scenarios for interleaved generation. We first leverage GPT-4o to generate a set of instances for each use case. For example, in marketing material generation, one instance is "*creating marketing campaigns around holidays to boost sales*". Then, we use GPT-4o to extend each instance into a more detailed instruction, e.g., "*Create an interleaved content that combines engaging text and eye-catching images for marketing campaigns around holidays to boost sales. Begin by researching holiday themes relevant to your products...*". Finally, we ask human annotators to verify whether the instructions are reasonable and of good quality. Note that we do not have gold references in this subset.

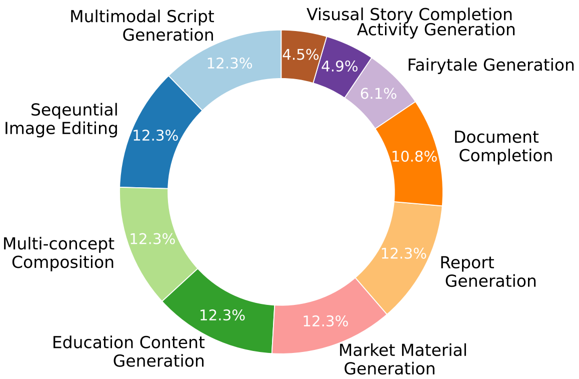


Figure 3: The distribution of the use cases in InterleavedBench.

Dataset Statistics In total, we finally collect 815 instances across 10 use cases, including *multimodal script generation*, *document completion*, *visual story completion*, *marketing material generation*, *report generation*, *education content generation*, *activity generation*, *sequential image editing*, and *multi-concept image composition*. The detailed distribution of the use cases is shown in Figure 3.

3.2 Comparison with Existing Benchmark

We highlight the following key differences and unique challenges introduced by our InterleavedBench compared with the existing benchmark. **(1): Output modality:** our benchmark requires the models to generate interleaved text and multiple images that could present in an arbitrary order, whereas exiting benchmarks (Kumari et al., 2023b) only cover the output with single modality or a single image (as shown in Figure 1); **(2) Requirement on coherence:** given that both inputs and outputs in our benchmark can contain multiple pieces of text and images, our dataset can assess whether the outputs are coherent and consistent with input instruction and context, and within the outputs themselves; **(3) Instruction following:** Most existing conditional image generation datasets only contain simple instructions such as "*add a cat next to the person*". On the contrary, each instance in our benchmark contains a detailed human-annotated instruction to describe the task. Thus, our dataset can evaluate models' instruction-following and generalization capabilities. We show the difference between our benchmark and existing datasets in Table 1.

4 InterleavedEval

In many use cases of interleaved generation, such as "*generate a story about Snow White using both text and images*", comparing the output against a gold reference is unrealistic since the generation can be fairly open-ended. However, existing approaches predominantly use reference-based metrics, e.g., BLEU (Papineni et al., 2002) and FID (Heusel et al., 2017), to measure the quality of text and image, respectively. They usually fail to assess the quality accurately.

To bridge the gap between existing metrics and the demand in more diverse and realistic scenarios, we present InterleavedEval, a strong reference-free metric based on GPT-4o, the current state-of-the-art LMM that supports arbitrarily interleaved inputs. To obtain a holistic and comprehensive evaluation of interleaved generation, we define five fine-grained evaluation aspects, including *text quality*, *perceptual quality*, *image coherence*, *text-image coherence* and *helpfulness*, and evaluate the output of each aspect separately. We show the detailed definition for each evaluation aspect in Table 5 in Appendix B. For each instance to be evaluated, the input of the evaluator consists of an **instruction** I that indicates what should be accomplished, **system output** $X = (T_O, \mathcal{P}_O)$, where T_O is the output text and $\mathcal{P}_O$ is the set of output images, the evaluation aspect a, and optionally, the **context** $\mathcal{C}$ of the task (e.g., the given text and images in models' inputs).

We formulate the evaluation metric InterleavedEval as follows: We instruct the GPT-4o evaluator to output discrete scores from {0, 1, 2, 3, 4, 5} based on the detailed criteria shown in Table 5, where 1 indicates the worst quality, 5 indicates the best quality, and 0 indicates output text and/or images are empty. We also instruct GPT-4o to provide a detailed explanation to improve the interpretability. Note that when the output text is empty, the scores on text-related aspects (*text quality* and *text-image quality*) are 0. Similarly, when the output image is empty, the scores on image-related aspects (*perceptual quality*, *image coherence*, and *text-image quality*) are 0. Moreover, we do not apply the text-related aspects in sequential editing and subject-driven generation since the primary focus of these tasks is whether the image is generated correctly according to the instructions.

5 Experiments

5.1 Experiment Setup

Baseline Models We benchmark the following baseline models which can be categorized into two types: *integrated models* where the LMM and image generation model are connected via neural modules, and *pipeline models* where the LMM and image generation model are connected via prompts in natural language. The integrated models include: **(1) MiniGPT-5 (Zheng et al., 2023a)** which connects a large language model with a stable diffusion model via generative vokens, enabling description-free multimodal generation; **(2) GILL (Koh et al., 2023)** which allows a pretrained large language model to generate multimodal responses by mapping the hidden states of text into the embedding space of an image generation model; **(3) EMU-2 (Sun et al., 2023a)** which induces in-context learning capabilities of LLMs by scaling up the model size and the size of the pretraining dataset; **(4) EMU-2 Gen + Gold Text** where EMU-2 Gen is a pretrained EMU-2 model instruction-tuned on various controllable image generation tasks. However, EMU-2 Gen cannot generate text so we combine it with ground-truth textual responses to come up with a complete text-and-image interleaved con-

Model	Text Quality	Perceptual Quality	Image Coherence	TIC	Helpfulness	AVG
MiniGPT-5	1.22	2.45	1.62	2.03	1.77	1.82
GILL	0.75	3.21	2.25	1.53	1.48	1.84
EMU-2	1.26	2.28	1.89	1.34	1.64	1.68
EMU-2 (Gold Text)	1.56	3.35	2.89	1.43	2.10	2.27
Gemini1.5 + SDXL	**4.40**	3.99	**3.64**	4.13	3.62	3.96
GPT-4o + DALL·E 3	4.37	**4.36**	3.51	**4.55**	**3.88**	**4.13**

Table 2: **Automatic evaluation** results of existing interleaved generation models on INTERLEAVEDBENCH using INTERLEAVEDEVAL. TIC is the abbreviation for 'Text-Image Coherence'. The best results are highlighted in **bold**.

Model	Text Quality	Perceptual Quality	Image Coherence	TIC	Helpfulness	AVG
GILL	1.35	1.89	1.72	1.43	1.19	1.52
EMU-2	1.23	1.74	1.87	1.24	1.2	1.46
Gemini1.5 + SDXL	**2.59**	2.36	**2.13**	**2.27**	2.08	2.28
GPT-4o + DALL·E 3	2.49	**2.51**	2.02	2.31	**2.13**	**2.29**

Table 3: **Human evaluation** results of existing interleaved generation models on INTERLEAVEDBENCH. TIC is the abbreviation for 'Text-Image Coherence'. The best results are highlighted in **bold**. Note that we use a scale of 0 to 3 for this evaluation, which is different from the scale used in Table 2.

tent for evaluation. The pipeline models include: **(5) GPT-4o (OpenAI, 2024) + DALL·E 3 (Betker et al.)** where GPT-4o is the state-of-the-art proprietary LMM that can comprehend interleaved text-and-image inputs and generate text-only responses. We leverage GPT-4o to generate text responses as well as captions for image responses in the desired positions. Then the captions are fed into DALL·E 3 to generate images. Finally, we combine the text responses with generated images in their original orders; **(6) Gemini-1.5 (Anil et al., 2023) + SDXL (Podell et al., 2023)**: we build this baseline in a similar way as GPT-4o + DALL·E 3 but use Gemini-1.5 Pro as the LMM and Stable Diffusion XL Turbo as the image generation model.

Baseline Metrics We adopt the following metrics as baselines to validate the effectiveness of our INTERLEAVEDEVAL. **(1) BERTScore** is a reference-based metric for text evaluation. We apply BERTScore to compute the similarity between the text output and the reference in our dataset. We set the BERTScore to 0 if the text output is empty. **(2) CLIPScore** is originally a reference-free evaluation metric for image captioning, which computes the cosine similarity between the CLIP embeddings of a predicted caption and that of the input image. We adopt CLIPScore as two baselines: a reference-based metric to compute image-image similarity between predicted images and ground truth images in a pair-wise manner, and a reference-free metric to compute the text-image compatibility between the generated images and text. **(3) DreamSim** is a recently proposed model-based metric to measure perceptual similarity. Similar to image-image CLIPScore, we use DreamSim to compute the perceptual distance between predicted images and ground truth images in a pair-wise manner.

5.2 Main Results

We show the main results of using INTERLEAVEDEVAL to conduct the fine-grained evaluation for various baseline approaches on INTERLEAVEDBENCH in Table 2. The baselines in the upper part are the *integrated* and *open-sourced* models while the baselines in the lower part are the pipeline models where the LMMs are proprietary. From Table 2, we observe that: **First**, the pipeline models consistently outperform the integrated models on all evaluation aspects by a significant margin, where GPT-4o + DALL·E 3 achieves the best performance on *helpfulness* and the average score of all the aspects. **Second**, the pipeline models achieve significantly good performance on *text quality* since Gemini and GPT-4o have strong text generation capabilities. Also, the generated visual prompts are generally coherent with the text content and they are directly fed into the image generation model, so the performance on *text-image coherence* of pipeline models is also remarkable. **Third**, we observe that the common errors of integrated models include the output text and/or im-

Metric	Ref-free?	Text Quality	Perceptual Quality	Image Coherence	TIC	Helpfulness
BERTScore	✗	0.21	-	-	-	0.37
DreamSim	✗	-	0.02	0.1	-	0.06
Image-Image CLIPScore	✗	-	0.08	0.2	-	-0.01
Text-Image CLIPScore	✓	-	-	-	0.2	0.09
InterleavedEval	✓	**0.72**	**0.30**	**0.43**	**0.4**	**0.57**

Table 4: **Mete-evaluation on evaluation metrics** in terms of Spearman correlation between automatic evaluation results with human judgments. For baseline metrics, we only report the correlation on the corresponding aspects (e.g., BERTScore can correspond to *text quality*) as well as *helpfulness*.

ages being empty, in poor quality, or having severe duplication. This is probably due to their weak instruction-following abilities. **Fourth**, *image coherence* is the most challenging aspect for the pipeline models. This is because the image generation model cannot take the images in the input context or previously generated images as conditions. Thus, the generated images do not have strong coherence. We include more qualitative analysis to illustrate these observations in Section 6.

5.3 Human Evaluation

In addition to automatic evaluation, we also conduct an extensive human evaluation to benchmark the baselines and also provide a meta-evaluation on our InterleavedEval and other evaluation metrics by computing the correlation between automatic evaluation scores and human judgments.

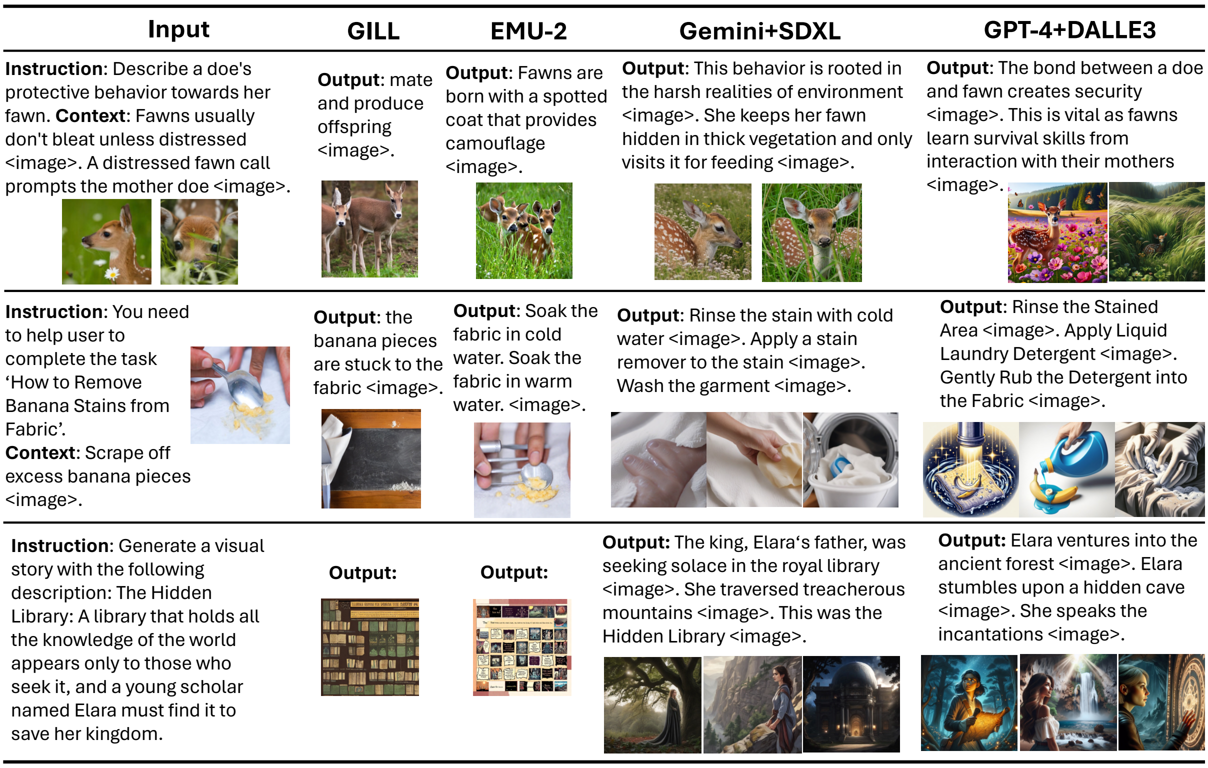
Human Evaluation Setup We adopt the same fine-grained evaluation criteria as InterleavedEval, where for each sample, the annotators need to give a score for each aspect defined in Table 5. The only difference is that, instead of rating on a scale of {0, 1, 2, 3, 4, 5}, we use a scale of {0, 1, 2, 3} for each aspect, where 1, 2, and 3 indicate the quality is *bad, fair*, and *good*, respectively. In this way, we can reduce the difficulty of human evaluation and improve its efficiency. Due to the cost of human evaluation, we select four representative baselines to evaluate, i.e., GILL, EMU-2, Gemini1.5 + SDXL, and GPT-4o + DALL·E 3. We include more details on human evaluation setup in Appendix B.1.

Results We show the human evaluation results in Table 3. The human evaluation is generally consistent with the automatic evaluation in Table 2. The pipeline models consistently outperform integrated models by a large margin, where GPT-4o+DALL·E 3 also achieves the best performance on *helpfulness* and the average performance. There's significant room for improvement in the integrated open-sourced models. We report the Inter Annotator Agreement (IAA) in Table 6 in Appendix B.1.

Correlation Analysis To validate the effectiveness of our proposed metric, we conduct a correlation analysis by comparing the evaluation results from automatic metrics with our human evaluation results. Since the baseline metrics only predict an overall score for each instance, we use the same set of evaluation scores to compare against the human rating on each aspect separately. For InterleavedEval, we compare evaluation scores with the human rating on corresponding aspects. Since most baselines require a gold reference, we use the context-based subset, where each instance has an associated reference output, to compute the correlation. From Table 4, our InterleavedEval consistently outperforms previous metrics by a significant margin in every aspect. Our metric has a particularly higher correlation on *text quality*, which is because *text quality* is relatively easier to evaluate with large language models like GPT-4o (Zheng et al., 2023b). Our metric achieves the lowest correlation on *perceptual quality*. The plausible reason is that GPT-4's perceptual recognition capability is still not strong enough to accurately detect visual artifacts or unnatural disruptions in the images (Fu et al., 2024). We also find that baseline metrics generally achieve poorer correlation, e.g., most metrics except for BERTScore almost do not have any correlation with *helpfulness*. BERTScore achieves the best correlation on *helpfulness* among baseline metrics, which indicates that text quality could be a good indicator of whether the overall interleaved content is helpful.

6 Discussions

Qualitative Analysis We conduct a qualitative analysis of benchmarked models in Figure 4 and have the following observations: (1) while GILL



Input	GILL	EMU-2	Gemini+SDXL	GPT-4+DALLE3
Instruction: Describe a doe's protective behavior towards her fawn. **Context**: Fawns usually don't bleat unless distressed <image>. A distressed fawn call prompts the mother doe <image>.	**Output**: mate and produce offspring <image>.	**Output**: Fawns are born with a spotted coat that provides camouflage <image>.	**Output**: This behavior is rooted in the harsh realities of environment <image>. She keeps her fawn hidden in thick vegetation and only visits it for feeding <image>.	**Output**: The bond between a doe and fawn creates security <image>. This is vital as fawns learn survival skills from interaction with their mothers <image>.
Instruction: You need to help user to complete the task 'How to Remove Banana Stains from Fabric'. **Context**: Scrape off excess banana pieces <image>.	**Output**: the banana pieces are stuck to the fabric <image>.	**Output**: Soak the fabric in cold water. Soak the fabric in warm water. <image>.	**Output**: Rinse the stain with cold water <image>. Apply a stain remover to the stain <image>. Wash the garment <image>.	**Output**: Rinse the Stained Area <image>. Apply Liquid Laundry Detergent <image>. Gently Rub the Detergent into the Fabric <image>.
Instruction: Generate a visual story with the following description: The Hidden Library: A library that holds all the knowledge of the world appears only to those who seek it, and a young scholar named Elara must find it to save her kingdom.	**Output**:	**Output**:	**Output:** The king, Elara's father, was seeking solace in the royal library <image>. She traversed treacherous mountains <image>. This was the Hidden Library <image>.	**Output:** Elara ventures into the ancient forest <image>. Elara stumbles upon a hidden cave <image>. She speaks the incantations <image>.

Figure 4: Case study. We select the representative examples of the system outputs from GILL, EMU-2, Gemini+SDXL, and GPT-4+DALLE3.

can generate images with reasonable quality, the generated text and images are typically not coherent with the instruction and context. In the example in the first row, the generated text is totally irrelevant to the task, while the image is also inconsistent with input images. (2) EMU-2 can often generate text that is relevant to the task, but the quality is not good enough. In the example in the second row, it repeatedly says "soak the fabric in water" but does not contain other useful content. Another weakness of EMU-2 is its poor conditional image generation capability, where generated images have obvious visual distortions and could be duplicated with input images. (3) On the other hand, the pipeline models can generally better follow the instructions and generate text and images in higher quality. Nevertheless, they still occasionally have some drawbacks. For Gemini+SDXL, some of the generated images (e.g., the first output image in the second example) still have obvious defects. For GPT-4+DALLE3, the style of generated images can be dramatically different from input images, as DALLE3 is prone to generate images in cartoon or dramatic styles. (4) Maintaining image coherence, i.e., the coherence of style and entities across images, is still very challenging for most models. In the third example, for the pipeline models, the same character has a very different appearance across the images, which makes the content inconsistent. (5) For the instances on the context-free subset, the integrated baselines have significantly worse performance, where they only generate one image with extremely poor quality. We hypothesize the reason to be those models cannot truly understand and follow the instructions. To sum up, our qualitative analysis indicates there is still significant room for improvement in interleaved generation.

7 Conclusion

We introduce INTERLEAVEDBENCH, the first benchmark for the evaluation of interleaved text-and-image generation. We also propose INTERLEAVEDEVAL, a strong multi-aspect reference-free evaluation metric based on GPT-4o. With extensive experiments, we first verify that our proposed metric can achieve significantly higher agreement with humans compared with existing metrics. Through the lens of INTERLEAVEDEVAL, we then observed that while the pipeline models based on proprietary LMMs consistently outperform open-source models, interleaved generation is still a challenging task that requires further advancement.

8 Limitation

While our proposed INTERLEAVEDBENCH and INTERLEAVEDEVAL provide a comprehensive evaluation suite for text-and-image interleaved generation, there are still several limitations in our work that we leave for future research. First, while INTERLEAVEDEVAL achieves the best correlation with human judgments among other evaluation metrics, it still does not have a high correlation on certain aspects, such as perceptual quality, image coherence, and text-image coherence. To further improve the evaluation accuracy, we may need to improve the capability of foundation multimodal models such that they are capable of recognizing subtle but critical differences. Second, our work did not extensively address the bias in using GPT-4 for evaluation, which we consider an important topic for future research.

References

Nantheera Anantrasirichai and David Bull. 2022. Artificial intelligence in the creative industries: a review. *Artificial intelligence review*, 55(1):589–656.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

James Betker, Gabriel Goh, Li Jing, † TimBrooks, Jianfeng Wang, Linjie Li, † LongOuyang, † JuntangZhuang, † JoyceLee, † YufeiGuo, † WesamManassra, † PrafullaDhariwal, † CaseyChu, † YunxinJiao, and Aditya Ramesh. Improving image generation with better captions.

Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. 2024. Subject-driven text-to-image generation via apprenticeship learning. *Advances in Neural Information Processing Systems*, 36.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *CoRR*, abs/2305.06500.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, Xiangyu Zhang, Kaisheng Ma, and Li Yi. 2024. Dream-LLM: Synergistic multimodal comprehension and creation. In *The Twelfth International Conference on Learning Representations*.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. BLINK: multimodal large language models can see but not perceive. *CoRR*, abs/2404.12390.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1233–1239.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. 2024. MANTIS: interleaved multi-image instruction tuning. *CoRR*, abs/2405.01483.

Jing Yu Koh, Daniel Fried, and Russ Salakhutdinov. 2023. Generating images with multimodal language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *International Conference on Computer Vision (ICCV)*.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023a. Multi-concept customization of text-to-image diffusion.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023b. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941.

Tianle Li, Max Ku, Cong Wei, and Wenhu Chen. 2023. Dreamedit: Subject-driven image editing. *Preprint*, arXiv:2306.12624.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *CoRR*, arXiv:2304.08485.

Stephanie M Lukin, Reginald Hobbs, and Clare R Voss. 2018. A pipeline for creative visual storytelling. *arXiv preprint arXiv:1807.08077*.

OpenAI. 2024. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`. Accessed: 2024-05-26.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. SDXL: improving latent diffusion models for high-resolution image synthesis. *CoRR*, abs/2307.01952.

Jingyuan Qi, Minqian Liu, Ying Shen, Zhiyang Xu, and Lifu Huang. 2024. Multiscript: Multimodal script learning for supporting open domain everyday tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):18888–18896.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. 2016. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2023a. Generative multimodal models are in-context learners. *CoRR*, abs/2312.13286.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2023b. Generative pretraining in multimodality. *CoRR*, abs/2307.05222.

Zhiyang Xu, Ying Shen, and Lifu Huang. 2023. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11445–11465. Association for Computational Linguistics.

Yue Yang, Artemis Panagopoulou, Qing Lyu, Li Zhang, Mark Yatskar, and Chris Callison-Burch. 2021. Visual goal-step inference using wikiHow. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2167–2179, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. 2023. Magicbrush: A manually annotated dataset for instruction-guided image editing. In *Advances in Neural Information Processing Systems*.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595.

Kaizhi Zheng, Xuehai He, and Xin Eric Wang. 2023a. Minigpt-5: Interleaved vision-and-language generation via generative vokens. *CoRR*, abs/2310.02239.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023b. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

A More Details on INTERLEAVEDBENCH

Data Filtering Pipeline To collect the source data from web resources, we first only keep the samples with 3 to 6 images and less than 12 sentences such that the ratio between text and image is balanced. We then apply Llama-8B-Instruct as a text filter to save the data with good text quality. We also apply LPIPS (Zhang et al., 2018) to discard the instances with duplicate images.

Manual Data Selection We apply a manual data selection and instruction annotation process to ensure data quality. We select the instances based on the criteria in Table 5. We also encourage the annotators to select diverse instances.

Instruction Annotation For each instance, we first ask an annotator to draft an instruction, and then ask another annotator to revise the instruction, until both annotators agree that the instructions are of high quality. The annotators are Ph.D. students with expertise in NLP and multimodal learning areas.

B More Details on Evaluation

We present the full list of our defined aspects and their definition in Table 5.

B.1 Human Evaluation

More Details on Human Evaluation Setup We sampled 100 instances from INTERLEAVEDBENCH as a subset for evaluation and ensure its task distribution is the same as the original distribution. In this way, we have 400 data points where each baseline has inference results on 100 instances. For each data point, we have two different annotators who are Ph.D. or master's students with expertise in NLP or multimodal domains to give ratings independently.

Inter-Annotator Agreement We show the IAA of our human evaluation in Table 6. While our human evaluation did not achieve significantly high agreement, we argue that the evaluation of interleaved generation is still quite subjective, open-ended, and challenging, even with our carefully designed human evaluation aspects and guidelines.

C Additional Experiment Results

Breakdown Results on Each Use Case We show a detailed breakdown of the average results on all the aspects of each use case. From Figure 5, we observe that (1) for pipeline-based models, image editing and subject-driven generation achieve the lowest results, whereas the models can achieve scores above 4 on other use cases; and (2) integrated models typically achieve low performance on the context-free subset in INTERLEAVEDBENCH. The potential reason is that these models did not specifically fine-turned on the data with text-only inputs, and thus cannot generate interleaved content well.

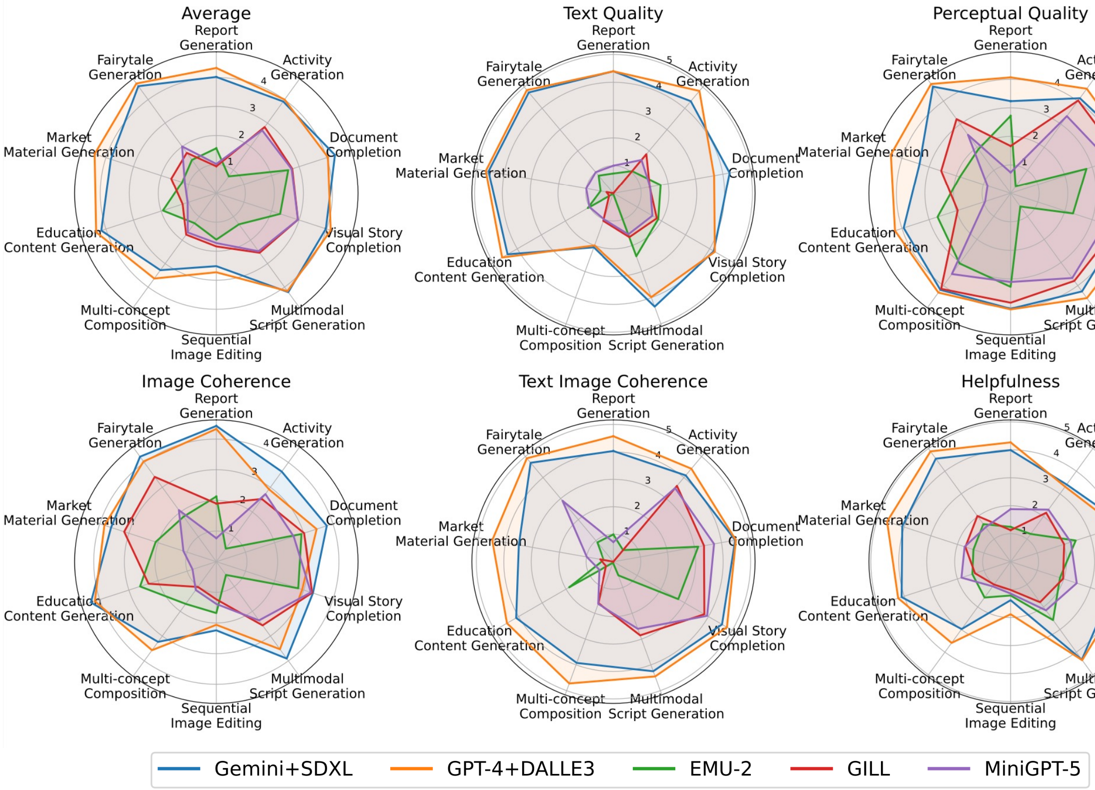


Figure 5: Breakdown performance on tasks.

Aspect	Definition
Text Quality	Text quality measures how clear, coherent, and error-free the output text is. It considers grammar, spelling, readability, coherence with the instruction and context, and whether it contains duplicate content.
Perceptual Quality	Perceptual quality measures how visually convincing, natural, and free from distortions or artifacts a generated image appears. It considers how accurately the image mimics reality without unnatural disruptions in structure, colors, or composition.
Image Coherence	Image coherence measures the consistency in style and subject representation across images. This includes textures, color palette, lighting, rendering styles, and maintaining consistent physical attributes, clothing, and behavioral traits. Image coherence also penalizes image duplication, where the output images are too similar, or within the output images themselves.
Text-Image Coherence	Text-to-image coherence measure the alignment and integration between textual and visual elements in a pairwise manner, ensuring they work together to convey a unified and cohesive narrative.
Helpfulness	Helpfulness measures how well the output text and images follow the task instructions and provide complete information to achieve the task. It also considers whether the outputs follow a reasonable logic flow.

Table 5: The full list of evaluation aspects and their corresponding definitions in InterleavedEval.

Text Quality	Perceptual Quality	Image Coherence	TIC	Helpfulness	AVG
0.489	0.306	0.320	0.427	0.519	0.412

Table 6: **Inter-Annotator Agreement** of human evaluation in terms of Cohen's Kappa score.