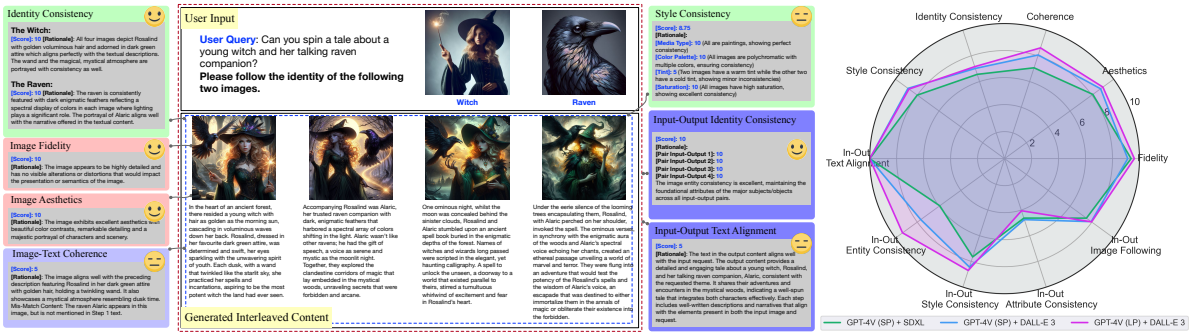
OpenLEAF: Benchmarking Open-Domain Interleaved Image-Text Generation

Anonymous ACL submission



(a) OpenLeaf's evaluation and justification on an interleaved sequence by our baseline method. (b) Performance comparison between baseline methods.

Figure 1: *OpenLEAF* can evaluate the quality of the interleaved content and give clear justifications like human (a), which is a robust tool to compare interleaved generative models' performance on diverse aspects (b). *SP* and *LP* denote short and long text-to-image prompts, respectively.

Abstract

We introduce *OpenLEAF*, a benchmark designed for the open-domain interleaved image-text generation task. This task aims to generate arbitrarily-interleaved multimodal content from input queries. It goes beyond single-modality image or text generation, thereby enabling various novel applications by creating content such as visual storybooks and how-to instructions. Despite the importance of the task, there lacks established benchmark due to the challenges in defining evaluation scenarios and formulating effective metrics. In this study, we collect a dataset covering queries with various input-output formats and 10 different application scenarios. We also propose an evaluation pipeline named "detection-summarization-scoring," which breaks down the evaluation into multiple reasoning steps. This pipeline leverages large multimodal models (LMMs) to thoroughly evaluate ten aspects of the generated content, which lead to the final rating. With experiments on a proposed agent system, we demonstrate that our evaluation method aligns closely with human judgments, offering a robust benchmark for assessing interleaved image-text generation.

1 Introduction

Open-domain interleaved image-text generation, aimed at generating multimodal content that follows user queries, has been desired for long in content generation. As shown in Fig. 1, the task is to generate an interleaved sequence of text descriptions and illustrative images that are coherent and correspond to user's text and image queries. The "open-domain" aspect indicates its adaptability to generate content freely across multiple domains, enabling the creation of complex content from visual instructions and illustrated stories to webpages. Such multimodal generation could redefine how we create and consume content in in various fields, such as education, marketing, etc.

Despite the significance of the task, there lacks standardized evaluation benchmarks, mainly due to the challenges in creating a comprehensive evaluation set and formulating effective metrics. In this study, we introduce *OpenLEAF*, a comprehensive benchmark that includes an evaluation dataset, a novel evaluation pipeline, and a baseline generation system, to standardize the task evaluation.

***OpenLEAF* Data and Topology.** One primary challenge of *OpenLEAF* is establishing standard evaluation queries and scenarios as a benchmark. With a focus on model generalizability and robustness (Zhang et al., 2021; Torralba and Efros, 2011), we categorize classes based on input formats and application domains and collect real user queries. This dataset includes 500 text-based

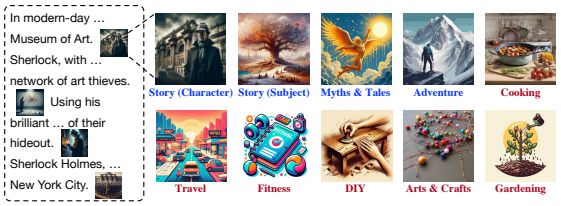


Figure 2: **Visualization of selected topics in *OpenLEAF*.** We show selected images in the generated interleaved content to visualize query scenarios in our *OpenLEAF* dataset. *OpenLEAF* covers diverse input-output formats and application scenarios. Blue image captions indicate the story generation task, while the red captions denote visual how-to instructions.

Evaluation Approach	Text Following	Image-Text Alignment	Image Attribution/Style Alignment
Chiang and Lee (2023)	✓	✗	✗
GPTEval (Liu et al., 2023c)	✓	✗	✗
GPTScore (Fu et al., 2023)	✓	✗	✗
CLIPscore (Radford et al., 2021)	✗	✓	✗
BLIP (Li et al., 2023c)	✗	✓	✗
VisualGPTScore (Lin et al., 2023b)	✗	✓	✗
Black et al. (2023)	✗	✓	✗
Liu et al. (2023b)	✗	✓	✗
LLMScore (Lu et al., 2023)	✗	✓	✗
X-IQE (Chen et al., 2023)	✗	✓	✗
Betti et al. (2023)	✗	✓	✗
***OpenLEAF* (Ours)**	✓	✓	✓

Table 1: **Capability comparison of open-domain evaluation methods in terms of the content alignment.** Thanks to the ability of pre-trained LMMs such as GPT-4V in capturing fine image details as well as a carefully designed prompting pipeline, *OpenLEAF* can evaluate complex mixed media alignments, facilitating the evaluation of interleaved content.

prompts across 10 different topics. Examples of these topics, as depicted in Fig. 2, include instructional cooking recipes, storybooks about specific characters, arts and crafts, etc. Besides, we collect 100 queries with reference texts, and 60 queries with image-text sequences as the reference to comprehensively test a model's capabilities.

***OpenLEAF* Evaluation.** Beyond data topology, another challenge is formulating an effective automated evaluation method suitable for interleaved image-text content. This method should comprehensively evaluate different aspects of the generation content. Existing evaluation methods, such as Visual-Language (VL) models (Radford et al., 2021; Li et al., 2022, 2023c; Chen et al., 2023) and specialized tools (Li et al., 2019; Maharana et al., 2021), fall short in evaluating detailed aspects like identity and style. Tab. 1 summarizes the capability of existing evaluation approaches, where none of them can be directly applied in *OpenLEAF*.

Inspired by the capabilities of LMMs (OpenAI, 2023b,c; Google, 2023; Microsoft, 2023; Yang et al., 2023), we use LMMs to evaluate open-domain content from various aspects of interest. We present a GPT-4V-based "detection-summarization-scoring" pipeline that comprehensively evaluates 10 aspects of the generated contents, i.e., *image fidelity, aesthetics, image-text coherence, identity consistency, style consistency, and the multi-modal input-output consistency in terms of the text, identity, visual style, subject's attribute, and overall image following ability.* For each aspect, the pipeline first *detects* key elements for that aspect, *summarizes* its findings, and *scores* the generated content based on the summarization.

Baselines. Ensuring semantic and visual style alignments across image-image, image-text, and text-text combinations poses a significant challenge in interleaved generation. Naively combining text-to-image (T2I) and image-to-text generations fail to generate coherent content. We present an agent system as the baseline for *OpenLEAF*, leveraging LMM and T2I models. Specifically, we use GPT-4V to generate texts and T2I prompts. To enhance the image consistency of generated sequences, global entity contexts and unified style descriptions are added to all T2I prompts. Then T2I models like DALL-E 3 (OpenAI, 2023d) and Stable Diffusion XL (SDXL) (Podell et al., 2023) are adopted to convert T2I prompts into images and form the interleaved sequence. We validate our proposed evaluation pipeline on this agent system through extensive analyses. Results show a strong correlation with human ratings of the generated interleaved content, indicating the effectiveness of the *OpenLEAF* benchmark.

Our contributions are summarized as follows.

- We build dataset of 660 samples for evaluating open-domain interleaved image-text generation, covering diverse formats and scenarios.
- We propose an LMM-based evaluation pipeline, named "detection-summarization-scoring," which evaluates generated content across 10 aspects with detailed justifications and ratings.
- We present a baseline agent system, and evaluate its generation with both LMM and humans. The high correlation with human ratings show that our pipeline is effective in evaluating multimodal contents, serving as a proxy for human ratings.

Input Type	Topic	# Problem
User Query Only	Cooking and Recipes	50
	Travel and Places	50
	Fitness and Well-being	50
	Home Improvement and DIY	50
	Arts and Crafts	50
	Gardening and Plant Care	50
	Story Requests about Specific Characters	50
	Story Requests about General Subjects	50
	Myths, Legends, and Historical Tales	50
	Adventure and Exploration	50
User Query + Text Context	Animal Stories	50
	Brand Introductions	50
Input Type	**Input Image Function**	**# Problem**
User Query + Text Context + Image Context	An image indicating the identity	20
	Dual images indicating identities	10
	An image indicating the image style	10
	An image indicating the attribute	10
	An image indicating the first image	10
Overall		**660**

Table 2: **A summary to our collected benchmark dataset**. The dataset covers diverse topics while setting challenges in satisfying various requirements related to input images.

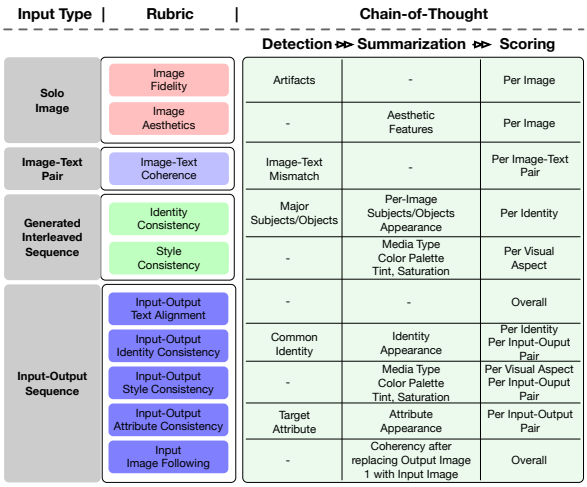


Figure 3: **An overview of the LMM-based evaluation pipeline.** We evaluate the quality of interleaved content on 10 aspects based on a detection-summarization-scoring process.

2 Related Work

Interleaved Image-Text Generation. Current interleaved image-text generation methods, including StoryGen (Liu et al., 2023a), AR-LDM (Pan et al., 2022), and StoryDALL-E (Maharana et al., 2022), primarily use fine-tuned latent diffusion models (LDMs) (Rombach et al., 2022) or text-to-image transformers. These approaches, however, are limited in open-domain generation due to fine-tuning on specific datasets. The challenge of evaluating open-domain interleaved content is still unresolved. Recent multi-modal LLMs like GILL (Koh et al., 2023), Emu (Sun et al., 2023), and DreamLLM (Dong et al., 2023) show promise in open-domain image-text tasks but are not specifically designed for interleaved generation and evaluation, leaving a gap in this area.

Foundation Models for Open-Domain Evaluation. Designing benchmarks (Trabucco et al., 2022; Zhu et al., 2023; Huang et al., 2023; Li et al., 2023a) and evaluating open-domain content has drawn increasing attention. In natural language processing, studies have shown the potential of prompting LLMs like GPT for open-ended text evaluation (Chiang and Lee, 2023; Liu et al., 2023c; Fu et al., 2023). For visual-language content, methods such as CLIPscore (Radford et al., 2021), BLIP (Li et al., 2022, 2023c), VisualGPTScore (Lin et al., 2023b), LLaVA-based scoring (Black et al., 2023; Liu et al., 2023b), LLMScore (Lu et al., 2023), X-IQE (Chen et al., 2023), and Betti et al. (2023) can effectively assess image-text similarity. However, they are limited to single image-text pairs and cannot fully evaluate interleaved content. Our work overcomes these limitations by using LMMs (OpenAI, 2023c) to evaluate open-domain interleaved content, accepting multiple image-text pairs for a more open-ended evaluation.

Multi-Modal Agents. As a baseline for open-domain interleaved generation, the presented agent system is related to multi-modal agent studies (Gupta and Kembhavi, 2023; Surís et al., 2023; Wu et al., 2023; Yang* et al., 2023; Shen et al., 2023; Li et al., 2023b; Lin et al., 2023a), which chain LLMs with multi-modal tools for new tasks. For example, Visual ChatGPT (Wu et al., 2023) shows that allocating various generative models (Rombach et al., 2022; Meng et al., 2021; Zhang and Agrawala, 2023) with ChatGPT (OpenAI, 2023a) can achieve complicated image generation and editing. Differently, our work focuses on a specific challenging task of open-domain interleaved image-text generation.

3 Method

This section details the evaluation dataset collection process, outlines our evaluation pipeline using LMM, and explains how we build the agent system using GPT-4V and T2I models to form open-domain interleaved generation baselines.

3.1 Data Collection

Tab. 2 outlines our evaluation dataset for open-domain interleaved content generation, detailing

problem types, formats, topics, and counts. The dataset is structured into three categories: user queries only, user queries with textual context, and user queries with multi-modal image-text context. In the user query only category, we assess the model's ability to generate varied topics like cooking recipes and DIY instructions. This includes 50 problems per topic, totaling 500 problems across 10 topics, created via GPT-4 prompts. The second category, encompassing user queries with textual context, features 100 problems where users add reference texts, such as stories or brand introductions, for integration into visual stories or webpage outputs. Lastly, the category involving user queries with multi-modal image-text context contains 60 problems. These problems are designed to assess the model's proficiency in following user-defined specifications, such as maintaining identity and visual style, aligning attributes with the input image, and generate interleaved content based on the input image. In sum, the *OpenLEAF* dataset offers a substantial set of 660 problems, providing a robust framework for evaluating performance in open-domain interleaved content generation.

3.2 Evaluation Pipeline

Our evaluation approach, utilizing GPT-4V (OpenAI, 2023c), imitates human judgment in analyzing interleaved content through a three-stage process: detection, summarization, and scoring. In the detection phase, GPT-4V examines the interleaved content to identify key elements relevant to our evaluation criteria. Next, in the summarization stage, it consolidates these findings into a clear summary, setting the stage for scoring. Finally, GPT-4V scores the content using a scale of $0, 5, 10$, where 10 represents excellent performance, 0 indicates significant flaws, and 5 reflects satisfactory content with minor issues. This streamlined method ensures a concise yet comprehensive evaluation. In our approach, as illustrated in Fig. 3, we assess the quality of interleaved content across 10 distinct aspects. These aspects are grouped into four categories, each reflecting a different scope of evaluation. The overall evaluation score is the average of scores obtained in each of these aspects.

Single Image. The evaluation of a single image encompasses both its fidelity and aesthetics. Fidelity assessment involves GPT-4V's analysis for artifacts such as blurriness or inconsistency, with scores given for visual fidelity. In aesthetics evaluation, GPT-4V scores the image based on artistic elements. These two aspects offering a rounded appraisal of an image's technical and artistic merit.

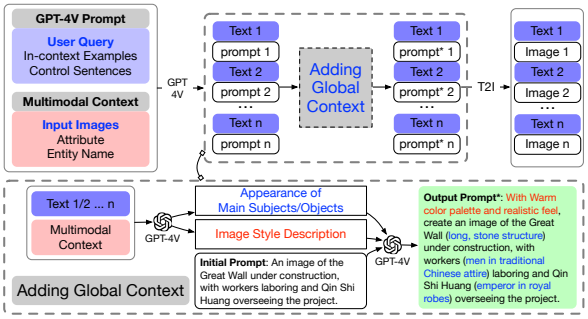


Figure 4: **A baseline interleaved generation method based on prompting LMM and T2I models**. We add global context into T2I prompts to encourage the visual consistency of images in the interleaved content.

Image-Text Pair. To evaluate the coherence of image-text pairs, we employ GPT-4V. Initially, both the image and corresponding text are inputted into the model. GPT-4V then detects any incongruities, assigning scores reflecting the number and importance of these discrepancies.

Generated Interleaved Sequence. To evaluate identity and style consistency, GPT-4V first extracts key subjects/objects from text descriptions, summarizes their appearance across images, and scores their visual consistencies across images. For style, it examines four dimensions: media type (realistic, painting, cartoon), color palette (monochromic, polychromic), tint (warm, cold), and saturation (high, low), scoring each image's adherence to these style parameters.

Input-Output Sequence. we evaluate the coherence of input-output multi-modal sequences on text alignment, identity and style consistencies, attribute consistency, and image following performance. Text alignment is assessed by comparing input and output texts through GPT-4V to ensure they match user requests. Identity and style consistencies are evaluated by comparing each input-output image pair, similar to output-only sequences. Attribute consistency is determined by extracting target attributes from user inputs and evaluating how well they're reflected in each input-output pair. Image following examines the content coherence when the first output image is replaced with the input image. These evaluations are conducted only for sequences where specific attributes, identities, styles, or starting images are provided by the user.

Method	Single Image↑		Image-Text Pair↑	Output Interleaved↑		Input-Output Interleaved↑					Average↑
	Fidelity	Aesthetics	Coherence	Identity Consistency	Style Consistency	Text Alignment	Entity Consistency	Style Consistency	Attribute Consistency	Image Following	
GPT-4V (*SP*) + SDXL	9.38	8.09	7.07	6.56	8.00	10.0	5.94	7.65	4.63	7.5	7.48
GPT-4V (*SP*) + DALL-E 3	9.13	8.80	8.07	7.02	8.85	10.0	7.68	8.44	4.75	7.94	8.07
GPT-4V (*LP*) + DALL-E 3	9.62	9.01	8.62	7.16	8.76	10.0	9.35	8.71	4.06	7.97	8.33

Table 3: **The comparison of baseline results.** *SP* denotes "Short Prompts", where T2I prompts contains at most 77 tokens per the requirement of SDXL. *LP* denotes "Long Prompts", where the word length of T2I prompt is roughly capped by 150. The highest and the second highest figures are highlighted by green and orange backgrounds, respectively. The full score is 10.

3.3 Interleaved Generation Baselines

We introduce an agent system that combines GPT-4V (OpenAI, 2023c) with text-to-image (T2I) models like DALL-E 3 or SDXL through textual prompts, setting a baseline for the open-domain interleaved generation task. As depicted in Fig. 4, the system initiates with a user query and a multi-modal input. A meticulously devised composition strategy crafts an input prompt detailing the desired content, format, contextual examples, and output constraints. These examples enable GPT-4V to grasp the expected content, fostering generation in the example's format and facilitating the automatic result extraction. Following this, GPT-4V processes the prompt and context, generates text descriptions, determines image insertion points, and crafts T2I prompts for image creation.

To ensure consistency in entity depiction and stylistic elements, we enrich the visual prompts with global entity and style contexts. The global entity context is a set of concise descriptions of each key subject's appearance. For the global style context, GPT-4V crafts it by selecting a suitable visual style for the content, such as recommending a warm color palette and realistic depiction for the Great Wall, as shown in Fig. 4.

4 Experiment

In this section, we focus on introducing the implementation details (Sec. 4.1), demonstrating comparison results of baselines (Sec. 4.2) and validate the effectiveness of our proposed LMM-based evaluation methods by qualitative, quantitative, and fidelity analyses (Sec. 4.3-4.5).

4.1 Implementation Details

We access GPT-4 and GPT-4V via the API. For T2I generation, we use SDXL v1.0 model [1] and the official DALL-E 3 (OpenAI, 2023d) API of November 2023. We keep all hyper-parameters of the SDXL to their defaults. For DALL-E 3, we use its "hd" mode and set the image resolution to 1024×1024. For LMM-based evaluation pipeline, we reset the system prompt of the GPT-4V to let it play the role of the interleaved content evaluator with a hint about subsequent jobs.

[1] https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0

4.2 Baseline Results on *OpenLEAF*

Tab. 3 presents the evaluation results of the GPT-4V-based agent system and its variants on the *OpenLEAF* set. Per-aspect scores are the average over all samples. The scores for the input-output interleaved sequences are only evaluated on problems with the corresponding image-text context. The overall score, an average of all 10 aspect scores, provides a comprehensive evaluation of the model's capability in the interleaved generation task.

In Tab. 3, we analyze three model variants that extend T2I prompt lengths and replace SDXL with DALL-E 3, aiming to examine how our evaluation pipeline can reflect these changes. Our analysis shows significant performance improvements of the agent system across most of the evaluated metrics with DALL-E 3, resonating the intuition that a better T2I model can lead to better interleaved results. Extending T2I prompts also improved scores in most aspects. However, longer T2I prompts did not enhance input-output attribute consistency. We hypothesize that more complicated T2I prompts might overload DALL-E 3's capability to accurately render attributes. These findings show that better T2I models and extended prompts effectively improve performance, validating the effectiveness of the *OpenLEAF* dataset and evaluation pipeline.

4.3 Qualitative Analysis

We conduct a qualitative comparison of interleaved content with high and low scores across each evaluation aspect. This analysis aims to determine whether our LMM-based evaluation pipeline effectively captures nuanced and comprehensive

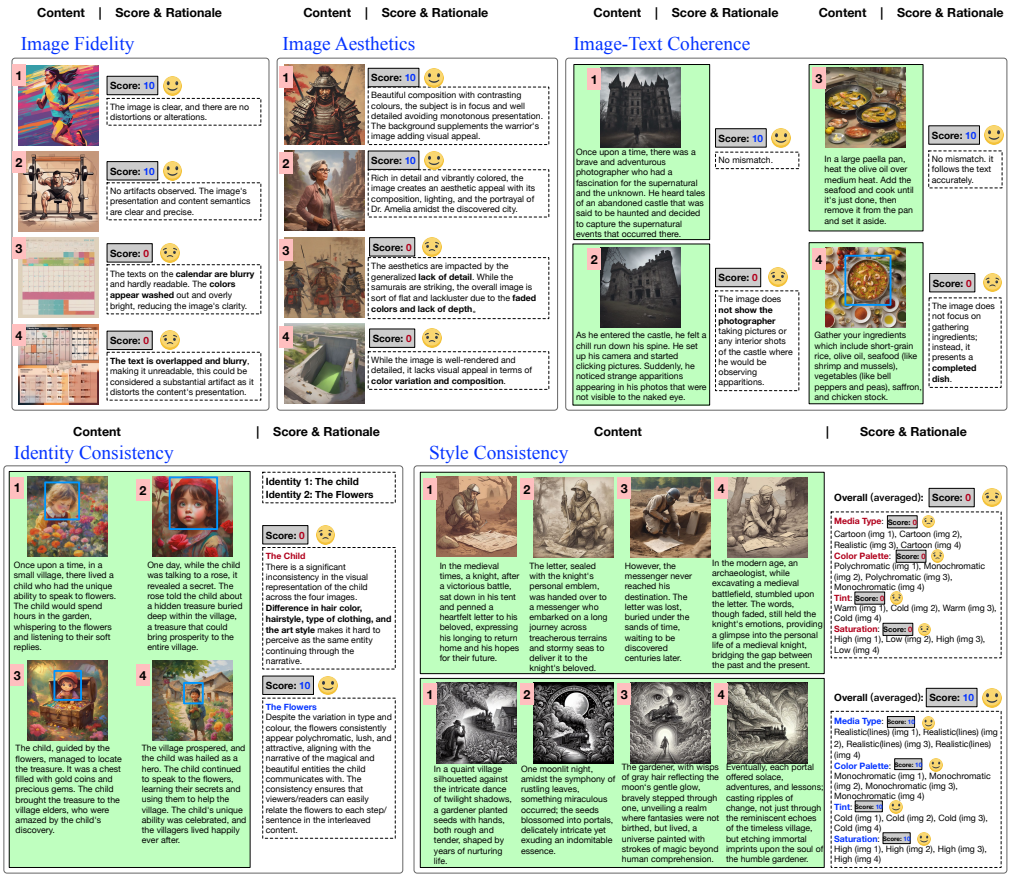


Figure 5: **The qualitative comparisons between the interleaved contents that yield high and low evaluation scores.** Blue boxes highlight image regions that cause low scores on an evaluation aspect. Our detection-summarization-scoring pipeline enables GPT-4V to understand both fine-grained and abstract image information, and give scores with logical rationales. Please zoom in to enjoy fine details.

multi-modal information, thereby providing scores backed by clear and logical reasoning.

Output Interleaved Sequence Evaluation. Fig. 5 shows qualitative comparisons of output interleaved content across various dimensions. Our method effectively spots image artifacts, including blurring and text rendering flaws. In terms of aesthetics, it identifies issues like faded colors and lack of depth in image 3, alongside poor color variation and composition in image 4. For image-text coherence, the approach excels in aligning objects and highlighting mismatches, such as the absent human figure in image 2 and the incorrect final dish in image 4. This precision extends to identity consistency, where it reliably recognizes main characters and their inconsistencies. For style consistency, *OpenLEAF* discerns stylistic elements, correctly categorizing images based on predefined visual styles, showcasing its good capabilities.

Input-Output Sequence Evaluation. Fig. 6 illustrates our method's effectiveness in evaluating input-output alignments for multi-modal queries and generated sequences. In assessing style consistency, it identifies variations in media types and notes the cold tint in image 4 of the top sequence. For identity consistency, it precisely spots age differences in the fisherman across input and output images, highlighting its meticulous attention to detail. Attribute consistency evaluation captures subtle changes in the style of goggles. The evaluation of image-following capability points out minor atmospheric variances between input and output images in the upper example. Additionally, the assessment of input-output text alignment showcases the system's reliability in generating accurate ratings, confirming GPT-4V's ability to produce coherent texts that follow user directives.

Collectively, the qualitative analyses in Fig.5 and Fig.6 demonstrate the *OpenLEAF*'s effectiveness in identifying both specific and broad aspects of multi-modal image-text content. *OpenLEAF* can conduct comprehensive and adaptable evaluation of interleaved content across different domains.

Figure 6: **The qualitative comparisons between the interleaved contents that yield high and low evaluation scores.** Our evaluation pipeline is effective in capturing high-level input-output connections. Please zoom in to enjoy fine details.

4.4 Quantitative Analysis

We conduct two types of quantitative evaluation to show the effectiveness of our LMM-based evaluation approach and analyze its behavior in comparison with humans. Our comparisons are based on 30 randomly selected interleaved content generated by our prompting-based interleaved generation baseline, which covers pure user query, user query with textual context, and image-text-conditioned generation scenarios. To collect human annotations, we conduct a user study and let human annotators give scores to each sequence for its image fidelity, image aesthetics, identity consistency, style consistency, and image-text alignment. We obtain 10 user responses in total for each sample. Next, we rank 30 samples on each aspect based on human scores, compared existing metrics (if available), and our evaluation scores. We report correlation scores of Kendall's τ and Spearman's ρ between the human rank and the rank by a compared evaluation method, indicating the similarity between the human annotation and the metric.

Tab. 4 compares our evaluation approach with established VL-based metrics: CLIP, BLIP-ITM, and BLIP-ITC on image-text alignment evaluation. CLIP has the worst performance among all the compared metrics. BLIP-ITM and BLIP-ITC have higher correlation scores and lower p values, indi-

Metrics	Spearman's ρ		Kendall's τ	
	Corr↑	p-value↓	Corr↑	p-value↓
CLIP (Radford et al., 2021)	0.2346	0.2206	0.1524	0.2576
BLIP-ITM (Li et al., 2022, 2023c)	0.3444	0.0673	0.2692	0.0455
BLIP-ITC (Li et al., 2022, 2023c)	0.3306	0.0799	0.2438	0.0701
Ours	0.5066	0.0043	0.3609	0.0047

Table 4: **The rank correlation comparison on image-text coherence between the established metrics and our evaluation approach.** The correlation scores are obtained based on image ranks from the human annotation and the compared metrics. The best and the second best results are highlighted by green and orange backgrounds, respectively.

Evaluation Aspect	Spearman's ρ		Kendall's τ	
	Corr↑	p-value↓	Corr↑	p-value↓
Image Fidelity	-0.0055	0.9767	0.2610	1.0000
Image Aesthetics	0.2610	0.1637	0.1862	0.1547
Image-Text Coherence	0.5066	0.0043	0.3609	0.0047
Identity Consistency	0.3295	0.0754	0.2460	0.0581
Style Consistency	0.4670	0.0093	0.2782	0.0314

Table 5: **The rank correlations between the human annotation and our evaluation approach on various aspects.** The human annotation is obtained by a user study consisting of 30 interleaved image-text samples.

cating their superiority over CLIP. Our evaluation approach achieves the highest correlation score and lowest p value, signifying its superior alignment with human evaluations in image-text alignment. For the remaining evaluation aspects, since no existing metrics can be directly applied, we show the correlation analysis between our LMM-based evaluation with the human below.

In Tab. 5, we utilize Kendall's τ and Spearman's ρ to show the rank correlation between our LMM-based evaluation and human annotations across five evaluation aspects. The table reveals a strong alignment between our method and human judgment in assessing image-text coherence and style consistency, as evidenced by high correlation scores and low p values. However, there is a notable divergence in the evaluation of image fidelity, where the LMM-based approach shows lesser agreement with human annotations, leading us to conduct a more in-depth analysis of image fidelity.

4.5 Fidelity Analysis

In Fig. 7, we showcase two interleaved samples to illustrate differences in image fidelity evaluations between GPT-4V and human judges. The top sample is rated as the lowest in image fidelity by GPT-4V but considered intermediate by humans. Conversely, the bottom sample is deemed the worst by humans but one of the best by GPT-4V. This discrepancy arises from the differing criteria used

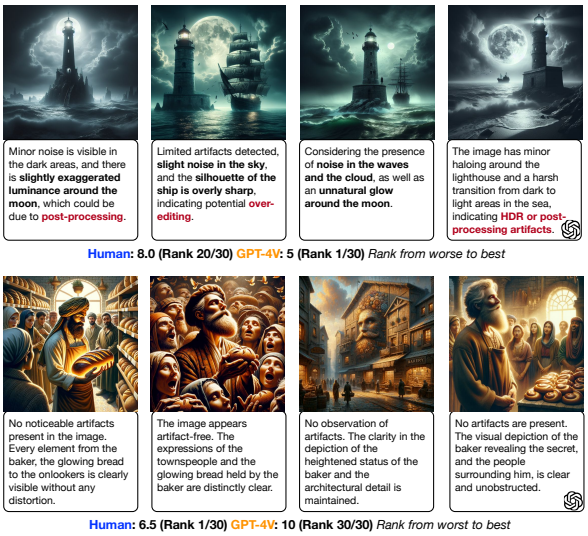


Figure 7: **The comparison between the contents favored by humans and GPT-4V on image fidelity.** GPT-4V pays more attention on pixel-space image noise, blurry, editing errors, and distortions. Oppositely, humans focus more on semantic-level image artifacts.

by GPT-4V and humans. GPT-4V concentrates on pixel-level details like noise, clarity, tone mapping errors, and gamma distortions. In contrast, humans focus more on semantic aspects of the images. For instance, in the bottom sample, humans notice semantic inconsistencies (e.g., an out-of-place face in the 3rd image), which GPT-4V overlooks. This divergence in focus and preference explains the low correlation between human and GPT-4V evaluations in terms of image fidelity. Recognizing that both perspectives offer valuable insights into image fidelity, we opt to retain the current fidelity evaluation prompt in our system and leave the human-aligned settings to be the future work.

5 Conclusion

In this paper, we focus on benchmarking the open-domain interleaved image-text generation by introducing an evaluation pipeline based on GPT-4V, a benchmark dataset to compare different approaches, and several baseline interleaved generation methods based on prompting GPT-4V and T2I models like DALL-E 3 and SDXL. Experimental results on the constructed benchmark dataset and a comprehensive analysis based on the user study demonstrates that the evaluation method based on GPT-4V can effectively evaluate the quality of an interleaved content in various aspects, where the evaluation pipeline, the benchmark dataset, and the baseline approach jointly form a reliable tool for benchmarking interleaved generative models.

6 Limitations

Unstable API Calls of GPT-4V. The presented baseline method involves utilizing the DALL-E 3 API by OPENAI to transform Text-to-Image (T2I) prompts into images. An observation is that the DALL-E 3 API occasionally fails to generate images for certain T2I prompts. This failure primarily comes from the API's internal safety checks. Consequently, our baseline model may not be able to generate interleaved content for certain queries.

Rating Variance. Another aspect worth discussing is the inherent variability in rating interleaved contents, especially for subjective aspects such as image fidelity and aesthetics. We aim at prompting GPT-4V to best mimic human evaluation processes when assessing interleaved content. As a result, its ratings also inherit a degree of variance akin to that of human evaluators. This variability positions GPT-4V's assessment more in line with subjective human judgment, rather than as a means of objective model-based evaluation.

7 Broader Impacts and Ethical Considerations

Broader Impacts. The introduction of *OpenLEAF*, a benchmark for the open-domain interleaved image-text generation task, represents a significant advancement in the field of multimodal content generation. By enabling the generation of content that seamlessly integrates both text and images, *OpenLEAF* paves the way for a myriad of novel applications. These include visual storybooks that can enhance learning and entertainment experiences, and detailed how-to instructions which could revolutionize education, technical manuals, and DIY guides by making them more accessible and engaging.

The ability to generate such integrated multimodal content has the potential to significantly impact industries such as education, publishing, and online content creation. For example, educational materials could be made more interactive and tailored to individual learning styles, potentially improving learning outcomes. In the realm of entertainment, personalized storybooks or interactive guides could offer unique experiences that engage users in unprecedented ways.

Ethical Considerations. While *OpenLEAF* focuses on the benchmarking aspect rather than the generation process itself, it inevitably raises ethical considerations related to the deployment and development of AI technologies capable of interleaved image-text generation, especially for the proposed interleaved generation baselines. One of the primary concerns revolves around the integrity and authenticity of content. As AI models become more capable at creating realistic and coherent multimodal content, distinguishing between AI-generated and genuine content becomes challenging, raising issues of misinformation and trust. The concern about the integrity and authenticity of AI-generated content is a common issue for all generative models.

Moreover, the benchmarking process itself must ensure fairness and avoid bias. The selection of datasets for evaluation and the definition of scoring criteria can inadvertently favor certain models or approaches. We have tried our best to ensure that *OpenLEAF*'s evaluation framework is designed to be as inclusive and unbiased as possible.

References

Federico Betti, Jacopo Staiano, Lorenzo Baraldi, Lorenzo Baraldi, Rita Cucchiara, and Nicu Sebe. 2023. Let's vice! mimicking human cognitive behavior in image generation evaluation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9306–9312.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. 2023. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*.

Yixiong Chen, Li Liu, and Chris Ding. 2023. X-iqe: explainable image quality evaluation for text-to-image generation with visual large language models. *arXiv preprint arXiv:2305.10843*.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. 2023. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Google. 2023. Bard. https://bard.google.com. Accessed: 2023-10-09.

Tanmay Gupta and Aniruddha Kembhavi. 2023. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pages 14953–14962.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *arXiv preprint arXiv:2307.06350*.

Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. 2023. Generating images with multimodal language models. *arXiv preprint arXiv:2305.17216*.

Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. 2023a. Agiqa-3k: An open database for ai-generated image quality assessment. *arXiv preprint arXiv:2306.04717*.

Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. 2023b. Multimodal foundation models: From specialists to general-purpose assistants. *arXiv preprint arXiv:2309.10020*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023c. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.

Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. 2019. Storygan: A sequential conditional gan for story visualization. In *CVPR*.

Kevin Lin, Faisal Ahmed, Linjie Li, Chung-Ching Lin, Ehsan Azarnasab, Zhengyuan Yang, Jianfeng Wang, Lin Liang, Zicheng Liu, Yumao Lu, et al. 2023a. Mm-vid: Advancing video understanding with gpt-4v (ision). *arXiv preprint arXiv:2310.19773*.

Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. 2023b. Visual-gptscore: Visio-linguistic reasoning with multi-modal generative pre-training scores. *arXiv preprint arXiv:2306.01879*.

Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, and Weidi Xie. 2023a. Intelligent grimm–open-ended visual storytelling via latent diffusion models. *arXiv preprint arXiv:2306.00973*.

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

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023c. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. 2023. Llmscore: Unveiling the power of large language models in text-to-image synthesis evaluation. *arXiv preprint arXiv:2305.11116*.

Adyasha Maharana, Darryl Hannan, and Mohit Bansal. 2021. Improving generation and evaluation of visual stories via semantic consistency. *arXiv preprint arXiv:2105.10026*.

Adyasha Maharana, Darryl Hannan, and Mohit Bansal. 2022. Storydall-e: Adapting pretrained text-to-image transformers for story continuation. In *ECCV*.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*.

Microsoft. 2023. Bingchat. https://www.microsoft.com/en-us/edge/features/bing-chat. Accessed: 2023-10-09.

OpenAI. 2023a. Chatgpt.

OpenAI. 2023b. Gpt-4 technical report.

OpenAI. 2023c. Gpt-4v(ision) system card.

OpenAI. 2023d. Improving image generation with better captions.

Xichen Pan, Pengda Qin, Yuhong Li, Hui Xue, and Wenhu Chen. 2022. Synthesizing coherent story with auto-regressive latent diffusion models. *arXiv preprint arXiv:2211.10950*.

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. *arXiv preprint arXiv:2307.01952*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*.

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

Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*.

Antonio Torralba and Alexei A Efros. 2011. Unbiased look at dataset bias. In *CVPR*.

Brandon Trabucco, Xinyang Geng, Aviral Kumar, and Sergey Levine. 2022. Design-bench: Benchmarks for data-driven offline model-based optimization. In *ICML*.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*.

Zhengyuan Yang*, Linjie Li*, Jianfeng Wang*, Kevin Lin*, Ehsan Azarnasab*, Faisal Ahmed*, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2021. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*.

Lvmin Zhang and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*.

Mingwei Zhu, Leigang Sha, Yu Shu, Kangjia Zhao, Tiancheng Zhao, and Jianwei Yin. 2023. Benchmarking sequential visual input reasoning and prediction in multimodal large language models. *arXiv preprint arXiv:2310.13473*.

A *OpenLEAF* Baseline Details

We present an agent system based on GPT-4V that can generate open-domain interleaved content in varying formats. Fig. 8 shows the results of our proposed baseline method in generating how-to instructions (top-left), visual stories (top-right), and converting textual content into visual stories (bottom-left) and webpages (bottom-right) via HTML and CSS codes. In the remaining part of this section, we show more details of the proposed interleaved generation baseline method.

The system achieves open-domain interleaved generation based on GPT-4V (OpenAI, 2023c) and T2I models like DALL-E 3 (OpenAI, 2023d) and SDXL (Podell et al., 2023). The top panel of Fig. 9 shows the generation pipeline of our method. Given an arbitrary user query, we initially follow a meticulously designed composition strategy to assemble an input prompt that indicates the content, format, and constraints of the target output. We then feed the input prompt and the optional image-text context into GPT-4V, which generates the textual descriptions, determines the positions to insert images, and formulates the visual prompt for each image. Subsequently, we incorporate global entity and style contexts into the visual prompts to improve the entity and style consistencies of SDXL. Here, the entity context comprises the appearance descriptions of common subjects, while the style context is a unique image style description shared across all visual prompts. Finally, T2I models convert visual prompts into real images, thereby creating the interleaved content.

Prompt Composition. The input prompts to GPT-4V consist of four parts. We first add a few in-context examples at the beginning. Each example shows the desired output corresponding to a specific input query. The in-context examples enable GPT-4V to comprehend the expected content more effectively and encourage it to generate content in the format of the in-context examples, facilitating easier automatic extraction of results. Subsequently, we concatenate the generation instruction with the user input to form the prompt. In this case, the instruction tells GPT-4V the desired output type, while the user input specifies the detailed content. Finally, we append control sentences to specify the number of image placeholders, story sentences, instruction steps, and `<div>`s in HTML, *etc*.

Text Generation. The first step of *OpenLEAF* baseline is to generate text. By feeding the assembled prompt discussed in the previous part, we enable GPT-4V to produce all text descriptions and image placeholders, indicating the position of each image. For example, as shown in the text generation panel of Fig. 9, when generating stories and how-to instructions, GPT-4V is prompted to generate story sentences and instructional steps, respectively, where image tags `<img{i}>` is also included in the generated text. Each image tag indicates the position of the corresponding image, forming an initial interleaved structure. When generating HTML code, the position of each image is determined by the placement of the `<img>` environment, where the generated CSS code can further tune the size, position, and alignment of each image. Next, we prompt GPT-4V to generate visual prompts from text descriptions. In this step, the input prompt also follows the composition strategy introduced earlier, incorporating all story sentences or instructional steps into the user input part. This approach allows GPT-4V to capture the context of the whole story or how-to instructions when generating the visual prompt for each image.

Adding Global Context. To improve the entity and style consistencies of images within the interleaved content, we introduce global entity and style context into the visual prompts before feeding it into T2I models to generate images. For the global entity context, we add a short appearance description of each common subject to the visual prompts, where GPT-4V is used to extract common subjects from text content, generate appearance descriptions, and rewrite visual prompts. To improve the style consistency of images, we prompt GPT-4V to determine a proper visual style to depict the interleaved content, based on the generated text descriptions. For example, GPT-4V indicates that a vibrant color palette and comic book style are best suited to illustrate superhero stories. Then, a short image style description is added to the beginning of each visual prompt to control the artistic style of images generated by T2I models. Fig. 10 compares the images generated from T2I prompts with and without using global context. The global context encourages the generated images to have better identity and style consistencies. The visual prompts equipped with the global context are then converted into images by T2I models, resulting in the interleaved content.

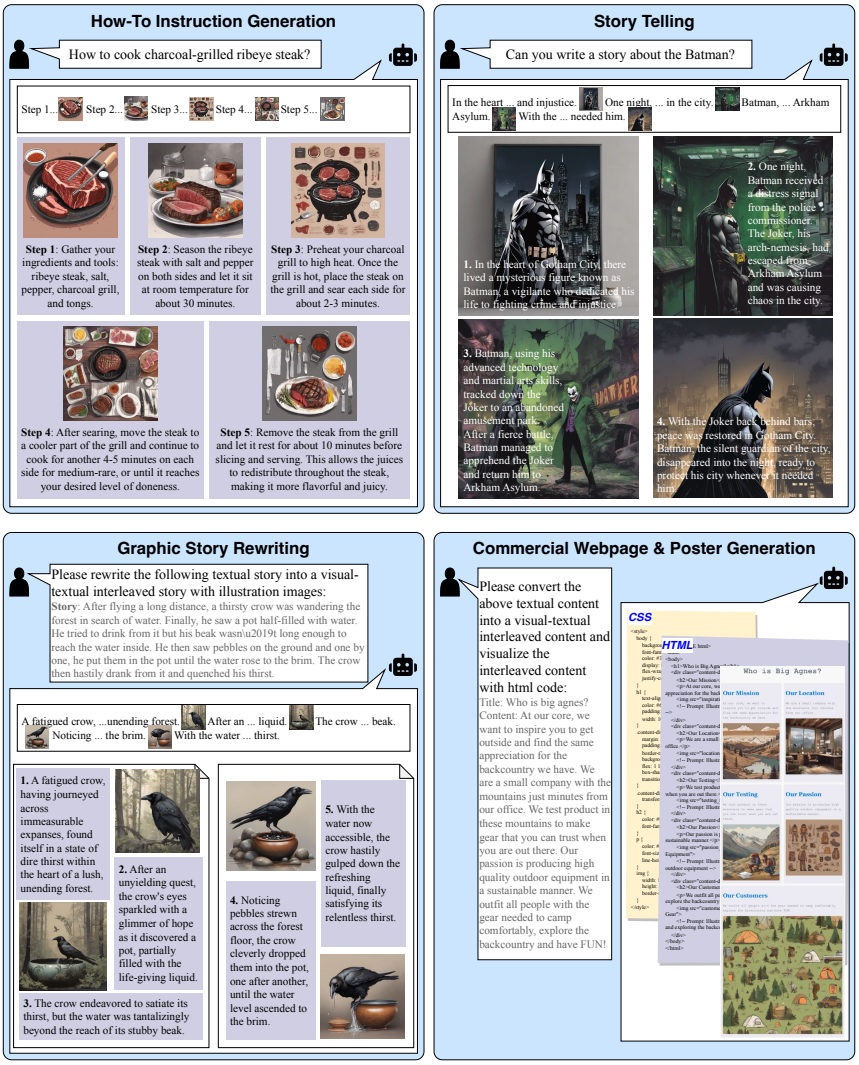


Figure 8: **The open-domain interleaved content generation results by *OpenLEAF* baseline method.** We show results on producing visual how-to instructions (top-left), generating multi-modal stories (top-right), converting textual stories to multi-modal stories (bottom-left), and generating webpages and posters via HTML and CSS codes (bottom-right).

B Z-Score Analysis

We conduct a Z-Score analysis to examine how closely the LMM-based evaluation mirrors human judgment and investigate GPT-4V's rating tendencies compared to human evaluators. We treat user ratings for each question as samples from an unknown distribution, estimating its mean and standard deviation (std) from these ratings. Then, we calculate the Z-Score for each GPT-4V evaluation and display these scores' distributions for each evaluation aspect in Fig. 11. On every aspect, most Z-Score's absolute values range within the $[-\text{std}, \text{std}]$, suggesting that LMM-based scores align with human annotation distributions. However, we notice a slight positive bias in the Z-Score distributions for image fidelity, aesthetics, style consistency, and image-text alignment, indicating GPT-4V tends to rate these aspects higher than human evaluators. Conversely, the identity consistency aspect exhibits a higher std but a mean close to 0, suggesting that while GPT-4V's ratings align with human evaluations overall, there is greater variability in its scores.

C User Study Interface

In Fig. 12 and 13, we show the interface for our user study. We use the Google Forms to present results to users and collect their responses. The first page of the user study shows the guideline about how to give scores for interleaved results on each aspect. Then in other pages, users are given a set of interleaved image-text results and we let users

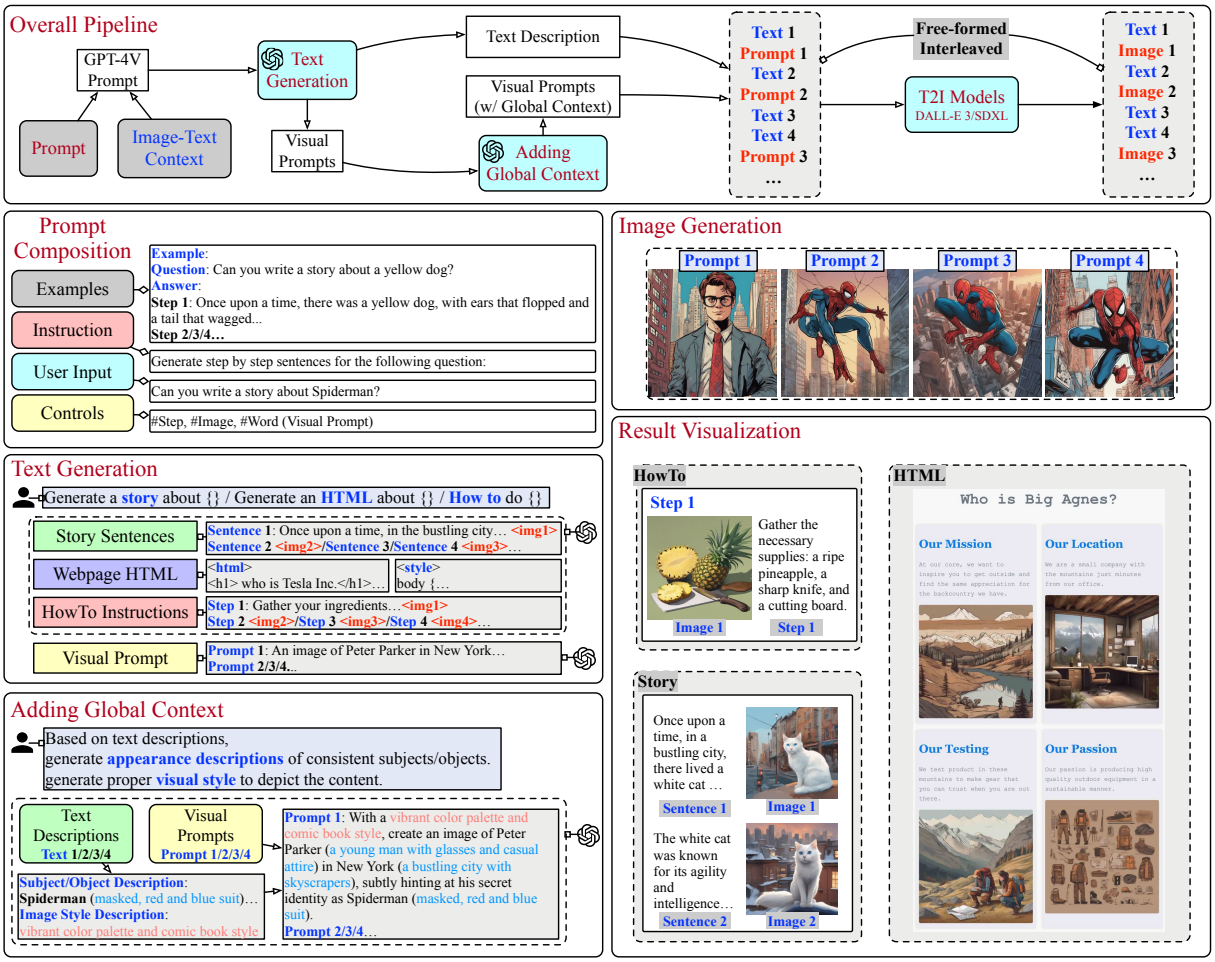


Figure 9: **Our interleaved generation framework.** The top panel illustrates the overall framework while other panels show details of each procedure.

to give score on each aspect based on the presented results.

D *OpenLEAF* Dataset Examples

The evaluation dataset of *OpenLEAF* contains 660 generation problems, which are categorized into three types according to the input format: user query only, user query with textual context, and user query with image-text context. In the following, we show a few questions for each input format and topic. The full evaluation set will be released together with the baseline and evaluation codes.

D.1 User Query Only

In this part, we show example problems in *OpenLEAF* evaluation set that only contain user queries in the input.

Cooking and Recipes (How-to) This group represents a category with a high frequency of real-world queries, is versatile, and is relatable for many people.

- `How do you make a classic French ratatouille?`
- `What are the steps to prepare and cook a beef Wellington?`
- `How can you make vegan chocolate chip cookies without eggs?`
- `What's the process to ferment your own sauerkraut at home?`
- `How do you prepare a traditional Spanish paella with seafood?`

Story Requests about Specific Characters (Story) This group captures the essence of story-telling with recognizable characters, allowing for the evaluation of creativity and fidelity to known character attributes.

- `Can you narrate a story where Sherlock Holmes solves a mystery in modern-day New York?`

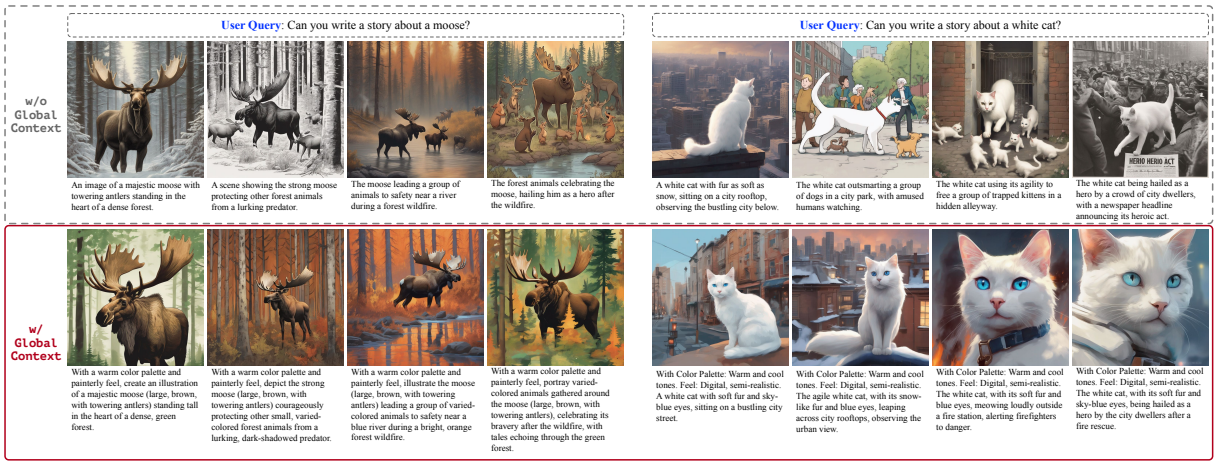


Figure 10: **The comparison between images generated from T2I prompts with and without using global context.** Using global context significantly improves the identity and style consistencies of images in the interleaved content.

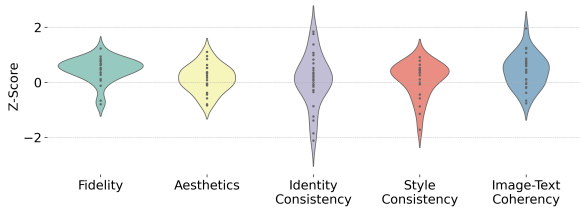


Figure 11: **The Z-Score analysis of *OpenLEAF* evaluation against human annotations.** *OpenLEAF* evaluation scores mostly fit the 1-σ range of the human's score distribution.

- How about a tale where Cinderella decides to start her own shoe business?
- What happens when Thor decides to take a break from being a god and lives as a common man on Earth?
- Tell a story about Snow White becoming an advocate for environmental conservation.
- How does James Bond handle an assignment when all his gadgets fail?

Travel and Places (How-to) This group represents a wide range of possible tasks, from planning trips to learning about world cultures, which is relevant to many users.

- How do you plan a two-week road trip across the U.S. Route 66?
- What are the top 5 recommended places to visit in Kyoto, Japan?
- How to pack efficiently for a month-long backpacking trip across Europe?
- What are the must-see landmarks when visiting Rome, Italy?
- How do you ensure safety while traveling solo in South America?

Fitness and Well-being (How-to) With an increasing focus on health and well-being globally, this category helps evaluate responses that touch upon both physical and mental well-being.

- How do you create a balanced weekly workout routine for weight loss?
- What are the steps to mastering the crow pose in yoga?
- How to develop a mindful meditation practice for stress relief?
- What's the best way to start a ketogenic diet for beginners?
- How do you build endurance for running a marathon?

Story Requests about General Subjects (Story) This group offers a more generic and flexible storytelling approach, allowing for the creation of new and diverse stories based on common subjects.

- Can you tell a story about a raindrop's journey through the water cycle?

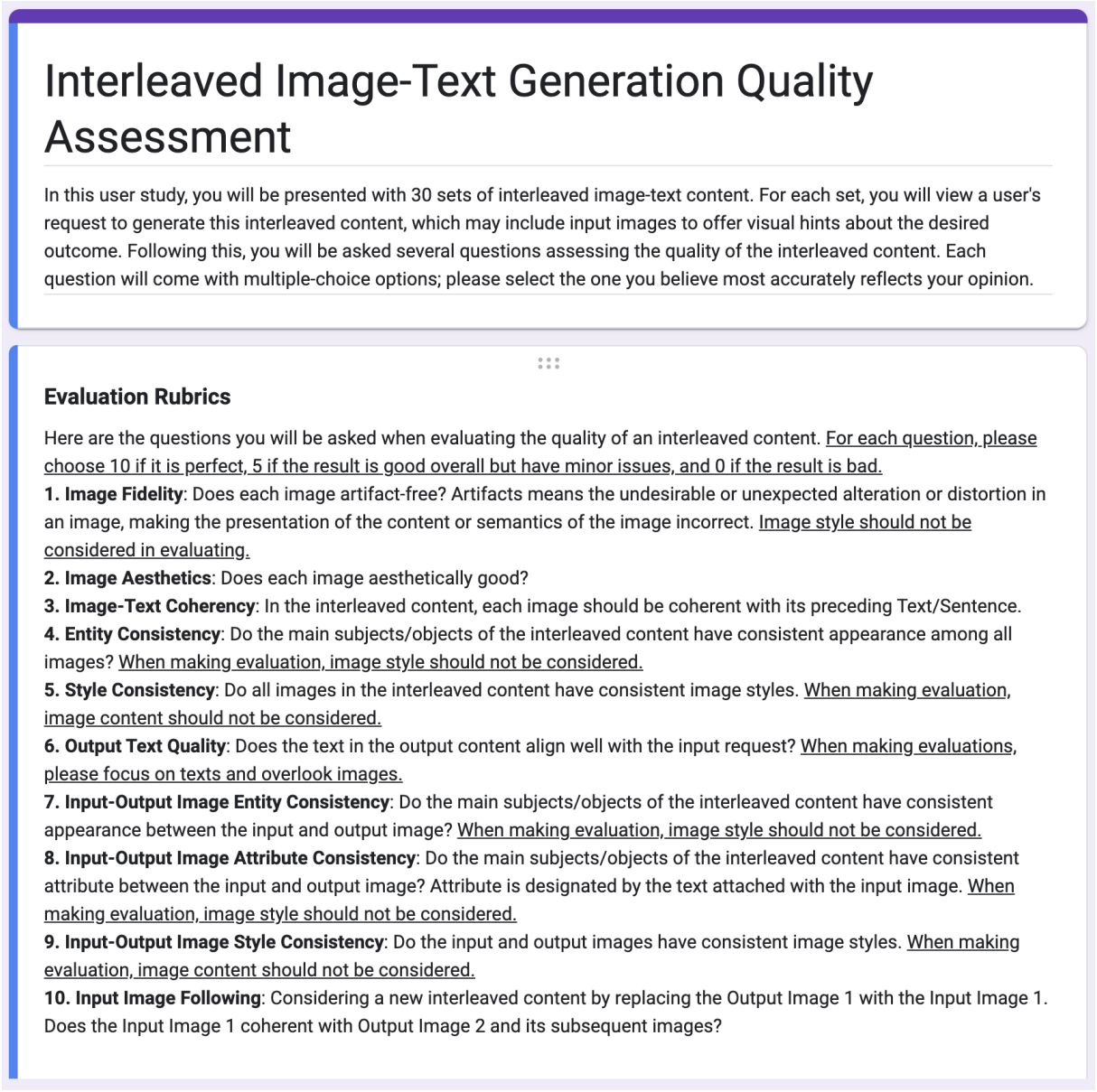


Figure 12: **The evaluating guideline of the user study.**

- Can you weave a tale about a tree that has witnessed centuries pass by?
- What's the story of a lonely lighthouse on a remote island?
- Can you narrate a tale of a mysterious mirror that shows one's true self?
- How about a story where the wind is a mischievous character playing tricks on a town?

Home Improvement and DIY (How-to) Given the prevalence of DIY and home improvement projects, this category tests practical advice and steps on common household projects.

- How do you install a floating shelf without visible brackets?
- What are the steps to refinish an old wooden table?
- How can you soundproof a room effectively for a home theater setup?
- What's the process to properly lay ceramic tiles in a bathroom?
- How do you create a vertical garden wall in your balcony?

Arts and Crafts (How-to) Creative in nature, this category evaluates the generation of artistic and craft-related content.

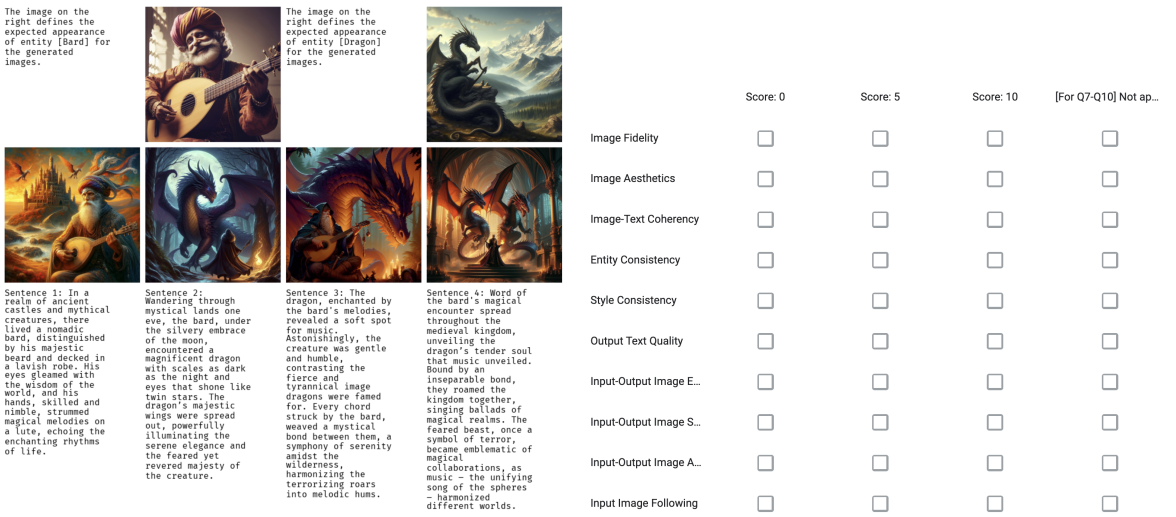


Figure 13: **The visualization of result and the evaluation panel in the user study.**

- How can you create a mosaic art piece using broken ceramics?
- What are the steps to knit a cozy winter scarf for beginners?
- How do you craft a homemade greeting card with 3D pop-up details?
- What's the process to sculpt a miniature clay figurine?
- How can you make DIY jewelry using beads and wires?

Myths, Legends, and Historical Tales (Story) This group focuses on retelling or crafting stories from the past, ensuring that the model's output adheres to known historical or mythical details.

- Can you recount the legend of King Arthur and the Knights of the Round Table?
- What is the story behind the Trojan Horse in ancient Troy?
- How did Icarus fly too close to the sun in Greek mythology?
- Can you narrate the tale of Cleopatra and her influence over Rome?
- What is the legend of the Lost City of Atlantis?

Gardening and Plant Care (How-to) This category can test for accuracy and depth of knowledge on a topic that many people are passionate about.

- How can you grow tomatoes in a container garden?
- What's the process for propagating succulents from leaves?
- How do you care for orchids to ensure they bloom annually?
- What are the steps to prepare garden beds for spring planting?
- How can you cultivate roses to produce vibrant blossoms?

Adventure and Exploration (Story) Pushing the boundaries of creativity, this category allows for the creation of new, exciting, and engaging stories centered around adventure themes.

- Can you weave a tale about an explorer discovering a hidden city beneath the Sahara Desert?
- What's the story of a young woman navigating the treacherous waters of the Amazon River?
- How did a group of adventurers uncover a secret entrance to Hollow Earth?
- Can you narrate a tale of an astronaut stranded on an alien planet and their journey back home?

- What's the chronicle of a mountaineer attempting to scale an unclimbed peak in the Himalayas?

D.2 User Query with Text Context

In this part, we show example problems in *OpenLEAF* evaluation set that contain textual context in the input.

Animal Stories

- In a verdant meadow, a rabbit often boasted about how fast it could run. Challenged by a tortoise, the overconfident rabbit took a nap midway through the race. The diligent tortoise, never pausing, continued its steady pace to cross the finish line first, proving that persistence often trumps speed.
- Perched high in a tree, a wise old owl observed the other birds chatter and squabble throughout the day. When asked why he was so silent, the owl replied, "The more I watch and listen, the more I learn. Sometimes silence teaches us more than noise."
- On the banks of a tranquil pond, a proud swan, mesmerized by its own reflection, dismissed the drab-looking ducklings nearby. However, as seasons changed, those ducklings matured into magnificent swans, illustrating that beauty often lies hidden beneath the surface.
- Deep in the jungle, a peacock displayed its vibrant feathers, catching the envy of other animals. Yet, when a heavy rainstorm hit, it was the humble sparrow, with its modest plumage, that found shelter and stayed dry, highlighting that outer beauty doesn't always provide inner strength.
- Near a bustling anthill, a grasshopper spent its days singing and dancing. While the ants toiled, storing food for winter, the grasshopper mocked their diligence. Yet, when winter came, the ants were well-prepared, and the grasshopper, regretting its laziness, went hungry.

Brand Introductions

- Title: Who is Nikon? Content: Nikon Corporation is a Japanese multinational corporation specializing in optics and imaging products. Founded in 1917, Nikon has established itself as a global leader in camera lenses, binoculars, and microscopes. Known for producing high-quality DSLRs, it has been a favorite among photographers for generations.
- Title: Who is Oracle? Content: Oracle Corporation is an American multinational computer technology company. Since its inception in 1977 by Larry Ellison, Bob Miner, and Ed Oates, it's provided software, cloud solutions, and hardware products, with its primary focus being databases. Oracle serves both the business and enterprise sectors with its vast range of services.
- Title: Who is LEGO? Content: LEGO Group is a Danish toy production company known globally for its iconic interlocking plastic bricks. Established in 1932 by Ole Kirk Christiansen, LEGO has expanded its realm, creating movies, games, competitions, and six themed amusement parks.
- Title: Who is Ubisoft? Content: Ubisoft Entertainment SA is a French video game company headquartered in Montreuil. Founded in 1986, it's now one of the largest in the industry. Renowned for franchises like Assassin's Creed, Far Cry, and Just Dance, Ubisoft continues to deliver immersive gaming experiences.
- Title: Who is Panasonic? Content: Panasonic Corporation is a Japanese multinational electronics company founded by Kōnosuke Matsushita in

1918. Initially a lightbulb socket manufacturer, it has grown into one of the largest Japanese electronics producers, alongside Sony, Hitachi, Toshiba, and Canon.

D.3 User Query with Image-Text Context

In this part, we show example problems in *OpenLEAF* evaluation set that contain image-text context in the input.

One Image Indicating the Identity This group contains one image per question that indicates the appearance of the main identity of the target interleaved content.

User Input: How to craft a handmade leather journal?

The image on the right defines the expected appearance of main entity for the generated images.

User Input: How to brew the perfect cup of herbal tea?

The image on the right defines the expected appearance of main entity for the generated images.

User Input: Can you write a story about a lonely lighthouse keeper and a mysterious ship that appears every night?

The image on the right defines the expected appearance of main entity for the generated images.

User Input: Can you craft a tale about a time-traveling librarian who seeks a forbidden book in ancient Alexandria?

The image on the right defines the expected appearance of main entity for the generated images.

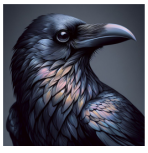
User Input: Can you narrate an adventure of a robot and a hummingbird who explore the ruins of a forgotten city?

The image on the right defines the expected appearance of main entity for the generated images.

Dual Image Indicating the Identities This group contains two images per question that indicating the appearance of the two main identities of the target interleaved content.

User Input: Can you spin a tale about a young witch and her talking raven companion?

The image on the right defines the expected appearance of entity [Raven] for the generated images.



The image on the right defines the expected appearance of entity [Witch] for the generated images.

User Input: Can you weave a narrative of a stranded astronaut and a benevolent alien on a distant planet?

The image on the right defines the expected appearance of entity [Alien] for the generated images.

The image on the right defines the expected appearance of entity [Astronaut] for the generated images.

User Input: Can you narrate a story of a treasure hunter and a mermaid exploring sunken ruins?

The image on the right defines the expected appearance of entity [Mermaid] for the generated images.

The image on the right defines the expected appearance of entity [Treasure Hunter] for the generated images.

User Input: Can you describe a tale of a nomadic bard and a dragon who loves music in a medieval kingdom?

The image on the right defines the expected appearance of entity [Bard] for the generated images.

The image on the right defines the expected appearance of entity [Dragon] for the generated images.

User Input: Can you detail a journey of a robotic engineer and a sentient AI navigating a post-apocalyptic world?

The image on the right defines the expected appearance of entity [AI] for the generated images.

The image on the right defines the expected appearance of entity [Engineer] for the generated images.

One Image Showing the Image Style This group contains one image per question that indicating the visual style of the target interleaved content.

User Input: Can you weave a story about a clockmaker who discovers a hidden world within one of his creations?

The image on the right defines the expected image style of the generated images.

User Input: Can you narrate a tale about a detective with the ability to hear the last words of the deceased?

The image on the right defines the expected image style of the generated images.

User Input: Can you recount an adventure about a pianist who can bring inanimate objects to life with her melodies?

The image on the right defines the expected image style of the generated images.

User Input: Can you describe a narrative involving a graffiti artist whose paintings predict future events?

The image on the right defines the expected image style of the generated images.

User Input: Can you craft a story about a gardener who plants seeds that grow into portals to other dimensions?

The image on the right defines the expected image style of the generated images.

One Image Showing the Attribute This group contains one image per question that indicating the appearance of a specific attribute of the target interleaved content.

User Input: Can you tell a tale about a mountaineer who finds a hidden city atop a peak?

The image on the right defines the expected appearance of attribute [Rugged outfit] for the generated images.

User Input: Can you weave a story of a potter whose clay sculptures come to life during a full moon?

The image on the right defines the expected appearance of attribute [Potter's wheel] for the generated images.

User Input: Can you narrate an account of a fisherwoman who captures a talking fish that offers three wishes?

The image on the right defines the expected appearance of attribute [Fisherwoman's hat] for the generated images.

User Input: Can you unfold a tale about a writer whose fictional characters start appearing in the real world?

The image on the right defines the expected appearance of attribute [Writer's quill] for the generated images.

User Input: Can you relay a narrative of a pilot who discovers a floating island in the sky?

The image on the right defines the expected appearance of attribute [Pilot's goggles] for the generated images.

One Image Showing the First Image This group contains one image per question that showing the first image of the target interleaved content.

User Input: Can you weave a tale about a librarian who discovers a book that writes its story in real-time?

The image on the right defines the expected first image for the generated images.

User Input: Can you describe a narrative involving a gardener whose plants hum tunes from old legends?

The image on the right defines the expected first image for the generated images.

User Input: Can you craft a story about a chef who can taste people's emotions in the ingredients they touch?

The image on the right defines the expected first image for the generated images.

User Input: Can you narrate a tale of a jeweler who creates amulets that grant the bearer's hidden desires?

The image on the right defines the expected first image for the generated images.

User Input: Can you recount an adventure of a tailor whose stitched clothes reveal the future of those who wear them?

The image on the right defines the expected first image for the generated images.

E Extra Interleaved Generation and Evaluation Results with the *OpenLEAF* Baseline

In this section, we first show two generation (Fig. 14 and 17) and raw evaluation outputs (Fig. 15, 16, 18, and 19) of our *OpenLEAF* baseline and evaluation approach, respectively. Taking Fig. 14 as an example, from the top to bottom, we show the user query at the top of the figure. Underneath it, we present the input images, if available, and state how the generated interleaved content should correlate with the input images. Next, we show all images and textual sentences from the generated interleaved content. For the raw evaluation outputs as shown in Fig. 15 and 16, we demonstrate the evaluation results including the score and the justification in each evaluation aspect, where the image quality includes the image fidelity and aesthetics. In each aspect, we follow the detection-summarization-scoring pipeline and show the evaluation output of each round one-by-one. We then show more interleaved generation results by the *OpenLEAF* baseline model in Fig. 20 to 29. We use DALL-E 3 for T2I generation in the interleaved content creation. The above figures demonstrate that *OpenLEAF* can produce high-quality interleaved content with coherent image-text sequence following the multi-modal context provided by users. The *OpenLEAF* evaluation outputs can capture fine-grained and abstract image details and give reasonable scores and clear justifications.

User Input: Can you conceive a narrative about an architect who designs buildings that heal the environment?

The image on the right defines the expected image style of the generated images.

Sentence 1: In a world suffused with hues of amber and azure, there lived an architect whose vision extended beyond concrete structures. Surrounded by landscapes treated as beautiful puzzles, the architect was inspired to breathe vivid life into his designs. Every blueprint he crafted possessed a symmetry and balance echoing nature's elegance.

Sentence 2: One day, amidst the harmonic dance of shadow and light, a revelation sparkled in the architect's mind. A concept of buildings become photosynthesizing entities, drawing toxins from the air and injecting oxygen back into the cities, where geometric lines met fluid grace.

Sentence 3: The architect had a gift for marrying straight angles with eloquent curves. His masterpiece: a structure where metal and glass spiraled upwards, mimicking the ascent of hot air balloons, breaking the monotony, and filling the skyline with colors reminiscent of a painting brushed by the fingers of twilight and dawn.

Sentence 4: As the transient sunlight glistened off the stained-glass extensions of the pinnacle, the architect's building stood as a testament to harmony. Between mankind and nature, it pulled the toxic whispers from the air and breathed out an aurora of healing echoes. A single edifice, yet it painted a future where architecture didn't just coexist but flourished midst nature's embrace.

Figure 14: **The generated interleaved content by the *OpenLEAF* baseline method.**

inout_text_quality
I would give it a score of 10. The text aligns well with the input request. Each step in the text provides a detailed progression of the architect conceptualizing, designing, and realizing buildings that contribute positively to the environment. The narrative flows coherently and is pertinent to the initial request about an architect who designs environmentally healing buildings.

---

inout_style_consistency
Round 1:

Aspect	Input Image 1 (painting, colorful)	Output Image 1 (painting, colorful)	Output Image 2 (painting, colorful)	Output Image 3 (painting, colorful)	Output Image 4 (painting, colorful)
Media Type (realistic, painting, cartoon)	painting	painting	painting	painting	painting
Color Palette (monochromatic, polychromatic)	polychromatic	polychromatic	polychromatic	polychromatic	polychromatic
Tint (warm, cold)	warm	warm	warm	warm	warm
Saturation (high, low)	high	high	high	high	high

Round 2:

Aspect	Input - Output Image 1	Input - Output Image 2	Input - Output Image 3	Input - Output Image 4
Media Type	10	10	10	10
Color Palette	10	10	10	10
Tint	10	10	10	10
Saturation	10	10	10	10

Round 3:

Calculating the average of each aspect:
1. Media Type: (10+10+10+10)/4 = 10
2. Color Palette: (10+10+10+10)/4 = 10
3. Tint: (10+10+10+10)/4 = 10
4. Saturation: (10+10+10+10)/4 = 10

Final score:
(10+10+10+10)/4 = 10

In conclusion, the style consistency score between the input image and each of the output images is a perfect 10. This indicates that there is excellent consistency in terms of media type, color palette, tint, and saturation across all the images presented. All images exhibit a painting style with high saturation, warm tint, and polychromatic color palettes, aligning perfectly with the style of the input image.

---

style_consistency
Round 1:

Aspect	Image 1	Image 2	Image 3	Image 4
Media Type	Painting	Painting	Realistic	Painting
Color Palette	Polychromatic	Polychromatic	Polychromatic	Polychromatic
Tint	Warm	Warm	Warm	Warm
Saturation	High	High	High	High

Round 2:

- Media Type Consistency: 5
- Color Palette Consistency: 10
- Tint Consistency: 10
- Saturation Consistency: 10

Explanation:

- For media type, there is a minor inconsistency: three images are paintings and one is more realistic.
- The color palette consistency is good across all images, they all are polychromatic.
- All images have a warm tint which is consistent throughout the interleaved content.
- Saturation is also consistent, with high saturation evident in all images.

Round 3:

Final Score = (5+10+10+10)/4 = 8.75

The style consistency score for the images in the interleaved content is 8.75. Despite minor inconsistencies in media type, the images maintain consistency in color palette, tint, and saturation, contributing to coherent visual aesthetics.

---

image_text_coherency
1. **Step 1 and Image 1:** The image fits the description perfectly. It portrays an architect surrounded by vivid, lively, abstract nature-inspired depictions. There are elements echoing the ambiance of a world "suffused with hues of amber and azure", and the architect appears to be inspired and in awe.
   - Mis-matched Content: None
   - Coherency Score: 10

2. **Step 2 and Image 2:** The text describes a revelation where buildings become entities drawing toxins and injecting oxygen back into the cities. The image captures a man who is presumably the architect, staring at a futuristic, vibrant cityscape that reflects the harmony of fluid grace and geometric lines. It visualizes the concept of nature-infused architecture but doesn't directly imply the buildings' photosynthesizing ability.
   - Mis-matched Content: The photosynthesizing effect of the buildings is not visually evident.
   - Coherency Score: 5 (The idea of nature-infused architecture is evident, though the specific element of photosynthesis isn't visually depicted.)

3. **Step 3 and Image 3:** In this pair, the text and image are in harmony with each other. The text speaks about a structure where metal and glass spiral upwards, with a color palette reminiscent of twilight and dawn. This imagery is explicitly and beautifully captured in the third image.
   - Mis-matched Content: None
   - Coherency Score: 10

4. **Step 4 and Image 4:** The text paints a picture of a building that stands as a testament to harmony, pulling toxic whispers from the air and breathing out healing, which is not visually evident. However, the image does illuminate the statement about the transient sunlight glistening off the glass, showing a building intricately, and beautifully integrated with nature.
   - Mis-matched Content: The building's ability to pull toxins and breathe out healing echoes is not depicted.
   - Coherency Score: 5 (The primary concept of harmony between the building and nature is evident, but specific details about its environmental healing properties are lacking visually.)

The overall content reflects minor mis-matches where conceptual descriptions in the text are not directly visually represented in the interleaved images, but the core essence of each step is largely captured, ensuring the expression remains coherent.

Figure 15: **The raw evaluation output for the generated interleaved content (Fig. 14) by the *OpenLEAF* baseline method.**

```
entity_consistency
Round 1:

The major subjects that are recognizable in the given interleaved content through the steps/sentences and images are the "architect" and the "architect's buildings with
nature-inspired designs."

Round 2:

1. **Architect:**
- **Image 1:** The architect is presented with a clear, detailed, and distinct appearance. He is shown holding blueprints, embodying the essence of his profession.
- **Image 2:** The entity consistency is moderately held, though the architect's features are not as distinct. The focus shifts towards the buildings and the fusion of
nature and architecture.
- **Image 3:** The architect is not visible. The focus is on the building manifesting as a result of the architect's innovative design vision.
- **Image 4:** Like Image 3, the architect is not seen. Concentration is on the portrayal of the building.

2. **Architect's Buildings with Nature-Inspired Designs:**
- **Image 1:** Nature's elegance is expressed using vivid, abstract art but no buildings are depicted.
- **Image 2:** The buildings become pronounced with nature-inspired designs, leading to a clear entity consistency with the textual content.
- **Image 3:** The building's design, inspired by nature, is prominent and aligns with the narrative.
- **Image 4:** The completion of the architect's vision is captured well, showing a building in harmony with nature.

Round 3:

1. **Architect:**
- Entity Consistency Score: 5
  - Explanation: There is a shift from a clear presentation of the architect to a more conceptual visualization, and in Images 3 and 4, the architect is not visible.

2. **Architect's Buildings with Nature-Inspired Designs:**
- Entity Consistency Score: 10
  - Explanation: Despite the gradual evolution from the conceptual introduction of nature-inspired designs to clear visual presentations in latter images, the theme remains
consistent.

Note: The scores are allocated considering the variations and consistencies in depicting these two primary entities. While the architect's physical presence diminishes, the
continuation of his vision through nature-inspired architectural designs is evident and consistent throughout the images.
------------------------------------------------------------
image_quality
### Task 1: Image Fidelity Evaluation

#### Image 1
- **Artifacts:** None noticeable
- **Score:** 10
- **Explanation:** The image appears clear without any noticeable artifacts or distortion, maintaining a high fidelity visual presentation.

#### Image 2
- **Artifacts:** None noticeable
- **Score:** 10
- **Explanation:** There is no visible alteration or distortion that negatively impacts the semantics of the image or its quality. The image expresses the content clearly.

#### Image 3
- **Artifacts:** None noticeable
- **Score:** 10
- **Explanation:** Image 3 is artifact-free, with seamless blendings of different elements, resulting in a high-quality depiction.

#### Image 4
- **Artifacts:** None noticeable
- **Score:** 10
- **Explanation:** Like the others, Image 4 also has a visual presentation. No significant artifacts can be identified and it aligns well with a high fidelity.

### Task 2: Image Aesthetics Evaluation

#### Image 1
- **Score:** 10
- **Explanation:** A visually captivating composition, characterized by exquisite detail, vibrant colors, and abstract nature-inspired elements which add a strong aesthetic
appeal.

#### Image 2
- **Score:** 10
- **Explanation:** This image offers a striking aesthetic appeal, showcasing a harmoniously intricate design, dazzling colors, and a beautifully illuminated cityscape.

#### Image 3
- **Score:** 10
- **Explanation:** With a breathtaking contrast of warm and dynamic colors amidst the serenity of the sky, and the innovative design of the building, this image possesses
exceptional aesthetic value.

#### Image 4
- **Score:** 10
- **Explanation:** The spectacular blend of colors reflecting off the building, coupled with the finely detailed architecture and radiant lighting, encapsulates a scene of
magnificent aesthetic beauty.

### Summary
The images are all artifact-free, offering clear, high-quality visual content. They hold significant aesthetic value characterized by vibrant colors, intricate designs, and
creative compositions that are both visually pleasing and emotionally evoking. The artistic and aesthetic synergy in the images exemplarily complements the textual context,
reinforcing the narrative flow and thematic coherence within the interleaved content.
------------------------------------------------------------
```

Figure 16: **The raw evaluation output for the generated interleaved content (Fig. 14) by the *OpenLEAF* baseline method.**

User Input: Can you craft a story about a chef who can taste people's emotions in the ingredients they touch?

The image on the right defines the expected first image for the generated images.

Sentence 1: In the bustling kitchen of a renowned restaurant, a talented chef with a meticulously groomed beard was known for creating unimaginable flavors. He discovered his extraordinary gift when he tasted a pinch of the sauce he was preparing and immediately experienced a wave of emotions.

Sentence 2: The chef realized that his unique ability allowed him to taste the emotions of the people who touched the ingredients. Each dish unveiled a story, a clandestine chapter of the lives intertwined with every flavor and aroma.

Sentence 3: Word spread about the empathetic chef, and people from far and wide visited his restaurant. They were curious and eager to taste dishes that mirrored their innermost emotions, eliciting tears of joy, laughter, and sometimes, silent reflections.

Sentence 4: The chef's extraordinary talent turned the restaurant into a sanctuary, where guests embarked on journeys of self-discovery through every bite. Food was no longer just a meal, but a profound experience of emotional revelation and unison of souls.

Figure 17: **The generated interleaved content by the *OpenLEAF* baseline method.**

```
inout_text_quality
I would give a score of 10, as the text in the output content aligns well with the input request. The output content adheres to the prompt by depicting a story about a chef
who can taste people's emotions through the ingredients they touch. Each step of the output content expands the narrative effectively, making it cohesive and directly
related to the request.
----------------------------------------------------------------
inout_following
To evaluate the potential replacement of Output Image 1 with Input Image 1, let's examine the coherence and quality of this new interleaved content.

Input Image 1 depicts a chef who appears to be in a state of surprise or curiosity as he tastes something, with a busy kitchen scene as the backdrop. The potential
incompatibility or coherence issues would arise depending on how well this image aligns with the accompanying text and subsequent images.

- Does the Input Image 1 coherent with Output Image 2 and its subsequent images?

In terms of the visuals, Input Image 1 aligns fairly well with Output Image 2, where the chef is depicted in a moment of realization about his unique ability to taste
emotions. The surprise and curiosity expressed in Input Image 1 could complement the dramatic explosion of flavors and emotions depicted in Output Image 2; the sequence of
emotions from curiosity to realization is visually logical.

On the continuity with subsequent images, Input Image 1 may slightly deviate. It effectively communicates the chef's astonishment or curiosity but might not carry the
otherworldly or transcendental quality that seems inherent in later images, especially Output Image 4. In this context, the original Output Image 1 does a better job of
introducing that mystical experience.

- Does the new interleaved content have good image-text coherency?

Considering the text, the replacement of Output Image 1 with Input Image 1 might not align perfectly with Step 1's description of the chef experiencing a wave of emotions.
Input Image 1's surprise expression doesn't necessarily communicate a profound emotional revelation the same way Output Image 1 does.

But, it could still work decently, as only a part of the emotional wave of experience would seemingly be missed. The bustling kitchen background in Input Image 1 certainly
complements the textual description of the kitchen scene.

Given these considerations:

- Input Image 1 can somewhat fit into the space of Output Image 1 in visual and emotional progression.
- The kitchen backdrop in Input Image 1 aligns well with the text.
- The emotional and mystical quality is slightly compromised.

I would assign a score of 5. The new interleaved content has minor issues, but it still maintains a good level of coherence for most audiences. The image can depict the
onset of the chef's realization of his gifts but may not convey the profound depth of emotions as effectively as the original Output Image 1.
----------------------------------------------------------------
style_consistency
### Round 1: Summarize the visual style of each image

Aspects |  Image 1 | Image 2 | Image 3 | Image 4
--- | --- | --- | --- | ---
**Media type (realistic, painting, cartoon)** | painting | painting | painting | painting
**Color Palette (monochromatic, polychromatic)** | polychromatic | polychromatic | polychromatic | polychromatic
**Tint (warm, cold)** | warm | warm | warm | warm
**Saturation (high, low)** | high | high | high | high

### Round 2: Score on the style consistencies of each aspect

Aspects | Score (0,5,10)
--- | ---
**Media type** | 10
**Color Palette** | 10
**Tint** | 10
**Saturation** | 10

Every image has been illustrated with a painting media-type, has a polychromatic color palette, holds a warm tint, and shares a high level of saturation.

### Round 3: Final Score of the Style Consistency

(10+10+10+10)/4 = **10**

### Explanation:

The interleaved content achieves a perfect score of 10 for style consistency. Each of the aspects, namely media type, color palette, tint, and saturation, are consistent
throughout the four images showcased in the content. Every illustration is a painting featuring a polychromatic color palette, warm tints, and high saturation. This
consistency in style allows for a cohesive and visually harmonious reader experience, enhancing the impact of the interleaved content.
----------------------------------------------------------------
image_text_coherency
### Image 1
#### Preceding Text/Sentence:
In the bustling kitchen of a renowned restaurant, a talented chef with a meticulously groomed beard was known for creating unimaginable flavors. He discovered his
extraordinary gift when he tasted a pinch of the sauce he was preparing and immediately experienced a wave of emotions.

#### Discovered Mis-matches:
No mis-matched elements found.

#### Score on Image-Text Coherency:
10

#### Explanation:
The image accurately reflects the chef in his busy kitchen tasting a pinch of the sauce and experiencing a wave of emotions. His astonishment matches the text description.

### Image 2
#### Preceding Text/Sentence:
The chef realized that his unique ability allowed him to taste the emotions of the people who touched the ingredients. Each dish unveiled a story, a clandestine chapter of
the lives intertwined with every flavor and aroma.

#### Discovered Mis-matches:
No mis-matched elements found.

#### Score on Image-Text Coherency:
10

#### Explanation:
The image represents the chef's realization of his unique ability vividly. The burst of colors and dynamic elements effectively portray the mix of emotions and stories
unveiled through every flavor and aroma.

### Image 3
#### Preceding Text/Sentence:
Word spread about the empathetic chef, and people from far and wide visited his restaurant. They were curious and eager to taste dishes that mirrored their innermost
emotions, eliciting tears of joy, laughter, and sometimes, silent reflections.

#### Discovered Mis-matches:
The image does not depict the visitors' reactions in terms of tears of joy, laughter, or silent reflections as expressed in the text.

#### Score on Image-Text Coherency:
5

#### Explanation:
The image depicts people gathering around the chef due to his reputation, in alignment with the text. However, it does not illustrate the various emotional reactions
elicited by the tasting of the dishes, as described in the text.

### Image 4
#### Preceding Text/Sentence:
The chef's extraordinary talent turned the restaurant into a sanctuary, where guests embarked on journeys of self-discovery through every bite. Food was no longer just a
meal, but a profound experience of emotional revelation and unison of souls.

#### Discovered Mis-matches:
No mis-matched elements found.

#### Score on Image-Text Coherency:
10

#### Explanation:
The image effectively captures the guests experiencing emotional revelations and unity through their meal, epitomizing the sanctuary-like ambiance of the restaurant due to
the chef's talent.

### Overall Assessment:
The coherency between the images and their preceding text is significantly strong. Three out of four images perfectly match the adjacent text, demonstrating an evident
correlation. One image has minor inconsistencies as it doesn't capture the range of emotional reactions described in the text but still aligns with the overall sentiment and
narrative. The seamless connection between the image and text iterations amplifies the potency and evocativeness of the interleaved content.
----------------------------------------------------------------
```

Figure 18: **The raw evaluation output for the generated interleaved content (Fig. 17) by the *OpenLEAF* baseline method.**

```
image_quality
### Task 1: Evaluate Image Fidelity

### Image 1:
#### Discovered Artifacts:
- No artifacts discovered.

#### Image Fidelity Score:
**10**

### Explanation:
The image is artifact-free. All elements in the scene are clear, distinct, and exhibit quality detailing, contributing to high-fidelity visual content.

### Image 2:
#### Discovered Artifacts:
- No artifacts discovered.

#### Image Fidelity Score:
**10**

### Explanation:
The visual content in Image 2 is also artifact-free, with clarity and precision in detail. The dynamic nature of the experience being portrayed is cleanly rendered.

### Image 3:
#### Discovered Artifacts:
- No artifacts discovered.

#### Image Fidelity Score:
**10**

### Explanation:
The quality is maintained in Image 3; character details, lighting, and surrounding elements are artifact-free, offering a clear and engaging visual presentation.

### Image 4:
#### Discovered Artifacts:
- No artifacts discovered.

#### Image Fidelity Score:
**10**

### Explanation:
Image 4 exhibits meticulous detailing of every character and object in the scene. The fluid aesthetic connect between various elements ensures that there are no unwanted
visual glitches or anomalies.

---

### Task 2: Evaluate Image Aesthetics

### Image 1:
#### Aesthetic Score:
**10**

### Explanation:
Image 1 offers a wonderful balance of colors, lighting, and emotions capturing the moment of astonishment of a chef with depth and expressiveness; it achieves high
aesthetic appeal.

### Image 2:
#### Aesthetic Score:
**10**

### Explanation:
This image stands out for its use of a vibrant mix of colors and dynamic illustrative patterns to paint a spectacle of abstract thematic revelation around the chef - it is
not just an image but an experience.

### Image 3:
#### Aesthetic Score:
**5**

### Explanation:
While there is clarity and detailing in Image 3, the aesthetic charm slightly pales in comparison to the previous images. The image feels like a passing sentiment with no
distinct aesthetic elements that arrest the viewer's eyes.

### Image 4:
#### Aesthetic Score:
**10**

### Explanation:
The final image rekindles the aesthetic journey with a blend of strategic lighting, color harmony, and intricate, detailed illustrations, creating a sensorially engaging
visual that elevates the diner's experience to a transcendental level.

---

------------------------------------------------------------
entity_consistency
### Round 1: Recognize Major Subjects/Objects
#### Major Subjects/Objects:
1. **The Chef:** The chef with a meticulously groomed beard is a central character in all the steps and should feature consistently across all images.
2. **The Emotions/Flavors:** While not a physical entity, the emotions and flavors brought about by the chef's cooking are another constant aspect depicted in the steps.

### Round 2: Summarize Key Features and Appearances
#### Major Subject/Object 1 - The Chef:
- **Image 1**: The chef has a long, groomed beard, and is wearing a chef's hat and uniform. He looks astonished.
- **Image 2**: The chef is still in his uniform and hat, with the same physical features including the beard.
- **Image 3**: The beard and attire remain consistent; the chef is presented in a posture of silent reflection or observation.
- **Image 4**: Physical features like the beard are still consistent with the previous images.

#### Major Subject/Object 2 - The Emotions/Flavors:
- **Image 1**: Emotions are captured through the chef's facial expression - a mix of surprise and exhilaration.
- **Image 2**: A burst of colors and dynamic elements depict an array of emotions and flavors exploding around the chef.
- **Image 3**: The focus shifts to the reactions of the people; however, their fixated gazes indicate strong emotions evoked by the chef's work.
- **Image 4**: The aura of the image portrays highly intense emotions; the guests' faces exhibit deep immersion and reflection.

### Round 3: Score on the Entity Consistencies
#### Major Subject/Object 1 - The Chef:
* **Score: 10**
* **Explanation:** The image of the chef is consistent across all steps. He is always depicted with similar physical features, particularly his signature groomed beard,
chef's uniform, and hat.

#### Major Subject/Object 2 - The Emotions/Flavors:
* **Score: 10**
* **Explanation:** The portrayal of emotions and flavors is also consistent. Each image captures either the chef's or the guests' emotional reactions to the food, or the
vibrant, abstract representation of flavors.

### Overall Entity Consistency Evaluation:
Both major subjects/objects, the chef and emotions/flavors enlightened by his cooking, are depicted with excellent consistency throughout the interleaved content, earning a
score of 10 for each entity. The chef's physical features and attire don't deviate, and there's a constant thematic portrayal of emotions and flavors, adhering to the
narrative seamlessly.
------------------------------------------------------------
```

Figure 19: **The raw evaluation output for the generated interleaved content (Fig. 17) by the *OpenLEAF* baseline method.**

User Input: Can you craft a story about a gardener who plants seeds that grow into portals to other dimensions?

The image on the right defines the expected image style of the generated images.

Sentence 1: In a quaint village silhouetted against the intricate dance of twilight shadows, a gardener planted seeds with hands, both rough and tender, shaped by years of nurturing life.

Sentence 2: One moonlit night, amidst the symphony of rustling leaves, something miraculous occurred; the seeds blossomed into portals, delicately intricate yet exuding an indomitable essence.

Sentence 3: The gardener, with wisps of gray hair reflecting the moon's gentle glow, bravely stepped through one, unveiling a realm where fantasies were not birthed, but lived, a universe painted with strokes of magic beyond human comprehension.

Sentence 4: Eventually, each portal offered solace, adventures, and lessons; casting ripples of change, not just through the reminiscent echoes of the timeless village, but etching immortal imprints upon the soul of the humble gardener.

Figure 20: **The generated interleaved content by the *OpenLEAF* baseline method.**

User Input: Can you craft a tale about a time-traveling librarian who seeks a forbidden book in ancient Alexandria?

The image on the right defines the expected appearance of main entity for the generated images.

Sentence 1: In a world overwhelmed by digital screens and artificial intelligence, a librarian with dark, short hair and clad in a long black coat, sought refuge in books and scrolls. Tucked within the anonymous corridors of a modern library, they discovered an ancient text that hinted at a forbidden book that possessed knowledge to transcend time and space.

Sentence 2: Consumed with curiosity, the librarian got hold of a peculiar metallic device inscribed with cryptic symbols from the realm of antiquities. As they adjusted the dials, a vortex of light spiraled, and they found themselves amidst the grandeur of ancient Alexandria, surrounded by intricate architecture and scrolls detailing the world's once-lost knowledge.

Sentence 3: The librarian, out of place amidst the grandeur of the ancient library, utilized their modern intellect and ancient texts, becoming engulfed in a cat-and-mouse chase through the towering shelves and architectural marvels of the space. The black coat swirled around them as they navigated the complexities of a civilization ahead of its time.

Sentence 4: As the distant roars of the ancient city echoed through the halls, the librarian with the dark, short hair clutched the forbidden book against the black fabric of their coat. Amidst the text lay secrets that could change the fate of timelines intersected. As they stared at the cryptic symbols found earlier, the vortex reappeared, paving the ascension to a world where past, present, and future became one.

Figure 21: **The generated interleaved content by the *OpenLEAF* baseline method.**

User Input: Can you depict a tale centered on a zookeeper who understands the secret language of animals?

The image on the right defines the expected first image for the generated images.

Sentence 1: In the tranquility of the night, under the dim lights, Mark, a dedicated zookeeper, found himself drawn to the majestic lion caged within the confines of the zoo. With his headphones on, Mark listened intently as the lion began communicating in a soft yet intense tone that only he could understand.

Sentence 2: Amazed and intrigued, Mark realized he had a unique gift - the ability to comprehend the enigmatic language of animals. Every whisper, grunt, and chirp disclosed a world full of untold stories and concealed emotions.

Sentence 3: As his connection with the animals deepened, the zoo transformed into a sanctuary where confessions of their past lives, hopes and dreams blossomed. Mark, engulfed with a newfound purpose, embarked on a mission to voice their silent pleas to the world.

Sentence 4: Mark's bond with the animals fueled a movement of transformation. The zoo reimagined, not as a place of captivity, but a haven of learning and coexistence. People far and wide, insufficient in understanding, came to witness the astonishing harmony between Mark and the zoo's residents. The animals, once imprisoned, now thrived as ambassadors of their kind.

Figure 22: **The generated interleaved content by the *OpenLEAF* baseline method.**

User Input: Can you describe a narrative involving a gardener whose plants hum tunes from old legends?

The image on the right defines the expected first image for the generated images.

Sentence 1: In the delicate embrace of dusk, where sunlight kissed the earth goodbye, a gardener tended to a mesmerizing garden adorned with flowers that painted the atmosphere with vibrant hues.

Sentence 2: Among the blossoming plants, the gardener noticed an enchanting hum. Each plant crooned tunes from old legends, weaving a melodic dance that swirled with the whispering winds.

Sentence 3: Bewitched, the gardener ventured deeper into the sonorous embrace of the garden each night, unraveling ancient tales spun by the symphony of flowering plants and blossoming trees.

Sentence 4: As the melodies seeped into the gardener's soul, they realized their garden was a sanctuary of eternal stories, where every petal and leaf was nurtured by the echoes of legends long past yet forever alive.

Figure 23: **The generated interleaved content by the *OpenLEAF* baseline method.**

User Input: Can you describe a narrative involving a graffiti artist whose paintings predict future events?

The image on the right defines the expected image style of the generated images.

Sentence 1: In a whimsical city where creativity knows no bounds, a graffiti artist named Alex explores the hidden alleys, painting walls with colors as vibrant and idealistic as a sunset behind a silhouette of enchanting woods.

Sentence 2: Alex discovers an enigmatic ability; every mural painted unravels a mystery of tomorrow, capturing scenes of future events touched with fantasy, as spirited and dreamy as a pond shimmering under the twilight sky, with frogs whimsically observing the glistening spectacle.

Sentence 3: The city begins to change; news of Alex's prophetic paintings spreads, and like the enchanting allure of nature's quiet places, garners an audience eager for hope, insight, or a glimpse into a tomorrow saturated with chars of otherworldly charm.

Sentence 4: Destiny awaits; just as the harmonious cycle of nature is inevitable, the narratives of Alex's graffiti stir a vibrant echo of curiosity, fear, and wonder in the city's heart, creating a world where urban life and mystical foresight intertwine under the enigmatic dance of day and night.

Figure 24: **The generated interleaved content by the *OpenLEAF* baseline method.**

User Input: Can you describe a story of a forest guardian and a lost unicorn seeking a magical grove?

The image on the right defines the expected appearance of entity [Guardian] for the generated images.

The image on the right defines the expected appearance of entity [Unicorn] for the generated images.

Sentence 1: In the depths of the silent forest, a mystical guardian veiled in a cloak of lush green leaves and piercing emerald eyes watches over. On an eventful day, the wind carries the soft whimpers of a lost unicorn, its coat as white as the first snow, glistened with a magical iridescence and a mane that danced with colors more vibrant than the most beautiful rainbow.

Sentence 2: The green-eyed guardian, touched by the unicorn's distress, steps out of the towering trees, the leaves in her attire rustling with the wind. She discovers that the magnificent unicorn is seeking the enchanted grove, a place rumored to possess the elixir of life, where the grass was greener, and flowers bloomed all year round.

Sentence 3: Guided by the guardian, they traverse the forest's depths where the starlight penetrates the thick canopy illuminating their path. The unicorn's radiant mane casts a technicolor glow, making the shadows and light perform a mystical dance around them. Despite the mesmerizing beauty, an air of urgency envelops them; for the magical creature is weakened, and the guardian's eyes reveal a steadfast determination.

Sentence 4: As dawn breaks, they reach the magical grove shrouded in an ethereal glow. The forest guardian with her green eyes gleaming, and the gleaming unicorn with its multicolored mane cascading, stand side by side witnessing the grove's magic unfold. With each step the unicorn takes, its strength rejuvenates, and the guardian's eyes reveal a mix of triumph and relief. Their alliance, forged in the pursuit, blooms - unveiling the arresting splendor and untapped powers of their mystical

Figure 25: **The generated interleaved content by the *OpenLEAF* baseline method.**

User Input: Can you describe a tale of a nomadic bard and a dragon who loves music in a medieval kingdom?

The image on the right defines the expected appearance of entity [Bard] for the generated images.

The image on the right defines the expected appearance of entity [Dragon] for the generated images.

Sentence 1: In a realm of ancient castles and mythical creatures, there lived a nomadic bard, distinguished by his majestic beard and decked in a lavish robe. His eyes gleamed with the wisdom of the world, and his hands, skilled and nimble, strummed magical melodies on a lute, echoing the enchanting rhythms of life.

Sentence 2: Wandering through mystical lands one eve, the bard, under the silvery embrace of the moon, encountered a magnificent dragon with scales as dark as the night and eyes that shone like twin stars. The dragon's majestic wings were spread out, powerfully illuminating the serene elegance and the feared yet revered majesty of the creature.

Sentence 3: The dragon, enchanted by the bard's melodies, revealed a soft spot for music. Astonishingly, the creature was gentle and humble, contrasting the fierce and tyrannical image dragons were famed for. Every chord struck by the bard, weaved a mystical bond between them, a symphony of serenity amidst the wilderness, harmonizing the terrorizing roars into melodic hums.

Sentence 4: Word of the bard's magical encounter spread throughout the medieval kingdom, unveiling the dragon's tender soul that music unveiled. Bound by an inseparable bond, they roamed the kingdom together, singing ballads of magical realms. The feared beast, once a symbol of terror, became emblematic of magical collaborations, as music – the unifying song of the spheres – harmonized different worlds.

Figure 26: **The generated interleaved content by the *OpenLEAF* baseline method.**

User Input: Can you design a story of a sailor who navigates using a compass that points to magical islands?

The image on the right defines the expected first image for the generated images.

Sentence 1: Under the ethereal glow of the full moon, an old sailor casts his gaze upon a mysterious compass in his hands, a ship anchored at a distance, heralding the beginning of an unforeseen odyssey.

Sentence 2: The compass, with powers unbeknownst to the sailor, didn't point north but to enigmatic islands where earthly logic held no sway; each direction leading to lands of distinct magic and allure.

Sentence 3: Trusting the enchanted compass, the sailor sailed through tumultuous seas and starlit nights, each island unveiling a world of mystique, with creatures and landscapes beyond human comprehension.

Sentence 4: The final destination marked the beginning of eternal legend; the old sailor, now a guardian of mystical realms, with his compass, not just a navigation tool, but a key unlocking the world's untethered imagination.

Figure 27: **The generated interleaved content by the *OpenLEAF* baseline method.**

User Input: Can you detail a saga of a baker whose bread can make anyone speak only the truth?

The image on the right defines the expected first image for the generated images.

Sentence 1: In a rustic bakery filled with the warm, enticing aroma of freshly baked bread, a rugged baker, well-known throughout the town, discovers his masterpiece - a unique loaf of bread that had an eye-catching golden brown crust and emitted a mystical glow, capturing the awe of onlookers.

Sentence 2: Word spread quickly about this magical bread; and curious, the townspeople couldn't help but take a bite. Astonishingly, anyone who tasted this delightful bread found that they could only speak the truth, their deepest secrets and most genuine thoughts flowing freely.

Sentence 3: As the truth-telling bread spread throughout the region, conflicts were resolved, relationships strengthened, and a newfound harmony enveloped the community. The baker became a celebrity, his bakery a pilgrimage for seekers of truth from distant lands.

Sentence 4: Bound by the spell of his own concoction, the baker revealed the secret of his enchanted bread. It wasn't the ingredients, but the pureness of his intentions and love in his craft that breathed life to the magic. Henceforth, he was revered as a sage; an ordinary baker with an extraordinary gift.

Figure 28: **The generated interleaved content by the *OpenLEAF* baseline method.**

User Input: Can you detail a story about a photographer who captures glimpses of the afterlife in his photos?

The image on the right defines the expected appearance of attribute [Photographer's camera] for the generated images.

Sentence 1: Alex, a revered photographer, was known for his unyielding fascination with the mysteries of life and death. Armed with his notable DSLR camera, every captured image told tales of the untold; yet the story to surpass them all was unfolding.

Sentence 2: During a midnight project, Alex decided to venture into an ancient, ominous forest. The moonlight filtered through the twisted branches created a divine dance of light and shadow. As he adjusted the focus on his DSLR, a chill ran down his spine.

Sentence 3: The first click of his camera that night captured more than just the eerie beauty of the forest, it arrested transitory glimpses of the afterlife. Ethereal beings, veiled in radiant gossamer threads of light, were articulating ancient tales of the beyond.

Sentence 4: Consumed by an insurmountable intrigue, Alex dedicated his life to understanding the gift his DSLR bore. Every snapshot unveiled a different perspective, a new narrative of the eternity. Mankind beheld the magnificent yet somber abyss of the afterlife through his lens.

Figure 29: **The generated interleaved content by the *OpenLEAF* baseline method.**