

# OPENLEAF: OPEN-DOMAIN INTERLEAVED IMAGE-TEXT GENERATION AND EVALUATION

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

This work investigates a challenging task named open-domain interleaved image-text generation, which generates interleaved texts and images following an input query. We propose a new interleaved generation framework based on prompting large-language models (LLMs) and pre-trained text-to-image (T2I) models, namely *OpenLEAF*. In *OpenLEAF*, the LLM generates textual descriptions, coordinates T2I models, creates visual prompts for generating images, and incorporates global contexts into the T2I models. This global context improves the entity and style consistencies of images in the interleaved generation. For model assessment, we first propose to use large multi-modal models (LMMs) to evaluate the entity and style consistencies of open-domain interleaved image-text sequences. According to the LMM evaluation on our constructed evaluation set, the proposed interleaved generation framework can generate high-quality image-text content for various domains and applications, such as how-to question answering, storytelling, graphical story rewriting, and webpage/poster generation tasks. Moreover, we validate the effectiveness of the proposed LMM evaluation technique with human assessment. We hope our proposed framework, benchmark, and LMM evaluation could help establish the intriguing interleaved image-text generation task.

## 1 INTRODUCTION

This work investigates an intriguing yet challenging task, namely open-domain interleaved image-text generation. As shown in Fig. 1, given an arbitrary open-domain query as the input, the task aims to generate a sequence of interleaved text descriptions and illustration images to form a coherent content following the input query. The ultimate goal of open-domain interleaved generation is to seamlessly generate arbitrary multi-modal contents, thus facilitating a wide range of applications and functionalities, such as generating multi-modal illustrations and web pages, storytelling, chain-of-thought explanations, and so on.

Early explorations (Li et al., 2019; Zeng et al., 2019; Li et al., 2020; Song et al., 2020; Maharana et al., 2021; Maharana & Bansal, 2021; Maharana et al., 2022; Szűcs & Al-Shouha, 2022; Pan et al., 2022; Liu et al., 2023a) simplify interleaved generation by narrowing down the problem to specific sub-domains, such as the story telling as shown in Fig. 1. Despite the promising explorations, these methods can only generate story illustrations in a format in which each image is paired with one single sentence. Therefore, these methods cannot achieve open-domain interleaved generation with arbitrarily interleaved content (Zhu et al., 2023; Gadre et al., 2023; Laurençon et al., 2023), where image and text are organized in diverse interleaved formats for input instructions covering a wide range of open-domain topics.

For the interleaved content evaluation, prior studies typically train a distinct evaluator for each aspect within a specific domain, such as a story character recognizer (Li et al., 2019; Maharana et al., 2021) for character consistency in story generation. However, these isolated evaluators cannot scale to open-domain interleaved generation due to the difficulty in generalizing to open-domain generative scenarios. Consequently, the remaining challenges of the open-domain interleaved generation are the lack of a proper evaluation approach for varying topics and diverse formats, a solid baseline, and a benchmark dataset to compare different methods on.

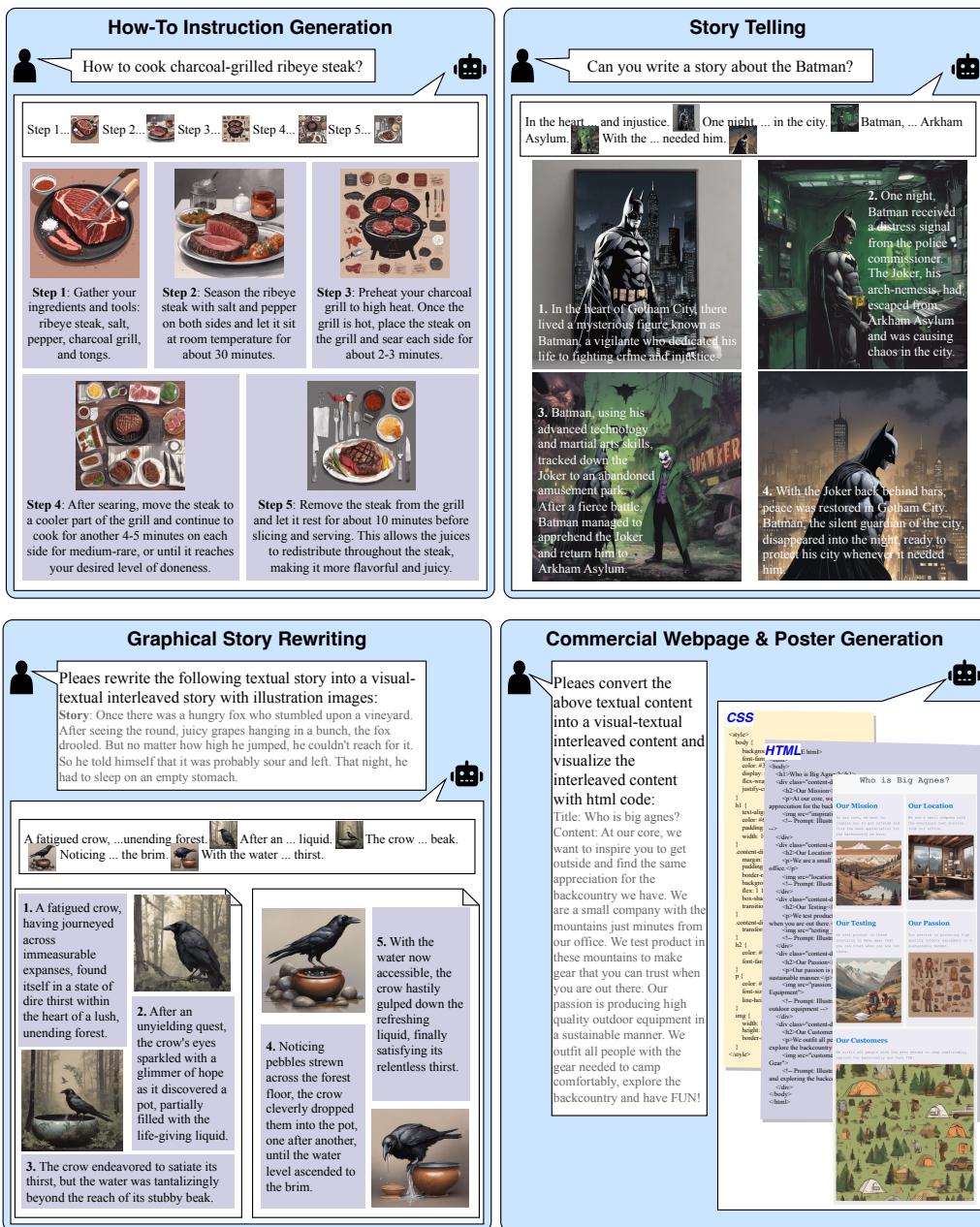


Figure 1: Examples of open-domain interleaved content generation. We show baseline results on producing visual how-to instructions (top-left), generating multi-modal stories (top-right), converting textual stories to multi-modal stories (bottom-left), and generating webpages and posters via HTML and CSS codes (bottom-right).

To address the lack of a baseline model, we propose a new training-free open-domain interleaved image-text generation framework based on GPT-4 (OpenAI, 2023b) and Stable Diffusion XL (SDXL) (Podell et al., 2023), named *OpenLEAF*. Given an arbitrary user query, *OpenLEAF* first prompts GPT-4 to generate a sequence of textual descriptions including image placeholders. Next we use GPT-4 to generate T2I prompts that fit well in the interleaved content. To improve the entity and style consistency of images within the interleaved content, we add global entity and style contexts to T2I prompts, where the entity context is short appearance descriptions of the common subjects throughout the interleaved content while the style context defines the target image style, both are produced by prompting GPT-4. Finally, SDXL converts T2I prompts into images, completing the interleaved content by replacing each image tag with the corresponding image.

For the interleaved content evaluation, inspired by the power of recent large multi-modal models (LMMs) (OpenAI, 2023b; Google, 2023; Microsoft, 2023), we explore prompting LMMs for unified interleaved content evaluation. Specifically, We use BingChat for the evaluation due to the empirical observation that it performs the best in assessing the quality of an interleaved image-text sequence from different perspectives. Ideally, an interleaved sequence should contain relevant and informative content, maintain consistency in subject identity and appearance, and feature a coherent image style. Therefore, we examine the entity and style consistency of the generated sequence, each broken down into sub-topics for LMM assessment. We then aggregate the generated scores as the final rating.

We collect a benchmark dataset for evaluating open-domain interleaved generation methods, which consists of 30 input queries, covering a wide range of topics and formats, such as visual instruction generation, story generation and rewriting, webpage, and poster generation. The experiments on this new benchmark based on the LMM-based evaluation approach demonstrate that the proposed baseline method can produce good interleaved results, achieving improvement against a simplified variant. In addition, we design comprehensive human evaluation and LMM analysis on the benchmark to validate the effectiveness of the BingChat evaluation.

Our contributions are summarized as follows.

- We study the open-domain interleaved image-text generation task, which aims to synthesize open-domain interleaved image-text sequences with arbitrary formats.
- We first explore LMM’s assessment ability to evaluate interleaved image-text sequences. Comprehensive human evaluation and analysis validate the effectiveness of the LMM-based evaluation technique.
- We propose *OpenLEAF* framework as a strong baseline for the interleaved image-text generation task, which improves the semantic consistency and style consistency with the proposed global context. We benchmark *OpenLEAF* on the developed evaluation set that covers diverse topics and sequence formats.

## 2 RELATED WORK

**Interleaved Image-Text Generation.** Starting from the GAN-based methods (Li et al., 2019; Zeng et al., 2019; Li et al., 2020; Song et al., 2020; Maharana et al., 2021; Maharana & Bansal, 2021; Szűcs & Al-Shouha, 2022), current methods that address the interleaved image-text generation task include StoryGen (Liu et al., 2023a), AR-LDM (Pan et al., 2022), and StoryDALL-E (Maharana et al., 2022), where Liu et al. (2023a) and Pan et al. (2022) fine-tune latent diffusion models (LDMs) (Rombach et al., 2022) on sequential story-like images, leading to auto-regressive LDMS, while Maharana et al. (2022) fine-tunes a text-to-image transformer. However, all of the above methods cannot be applied to the open-domain interleaved generation since the fine-tuning process makes the above model only work well on images similar to the fine-tuning datasets, leading to limited generation domains. On the other hand, how to evaluate open-domain interleaved content remains to be unsettled.

Recent multi-modal LLMs such as GILL (Koh et al., 2023), Emu (Sun et al., 2023), and Dream-LLM (Dong et al., 2023) show the decent performance on open-domain image-text generations and perceptions. However, they are not specifically designed for interleaved generation and evaluation, leaving the open-domain interleaved generation task unestablished.

**Foundation Models for Open-Domain Evaluation.** How to evaluate open-domain content has drawn increasing attention. In the natural language process, studies show promises of prompting LLMs such as GPT for open-ended text evaluation (Chiang & Lee, 2023; Liu et al., 2023c; Fu et al., 2023). For evaluating the visual-language content, CLIPscore (Radford et al., 2021), VisualGPTScore (Lin et al., 2023), and LLaVA-based scoring methods (Black et al., 2023; Liu et al., 2023b) effectively evaluates the open-domain image-text similarity. However, these methods can only work on single image-text pairs while having limited capabilities to evaluate arbitrarily interleaved content comprehensively.

This work uses LMMs (*i.e.*, BingChat (Microsoft, 2023)) for evaluating the open-domain interleaved content, which addresses the above-mentioned issue by accepting multiple image-text pairs and allowing for open-ended evaluation.

**Multi-Modal Agents.** The method part of this paper is related to multi-modal agent studies (Gupta & Kembhavi, 2023; Surís et al., 2023; Wu et al., 2023; Yang\* et al., 2023; Shen et al., 2023; Li et al., 2023), 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 & 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 introduces *OpenLEAF* and the evaluation pipeline based on LMM. Fig. 2 overviews our method. Subsequently, we detail our generation and evaluation pipelines, respectively.

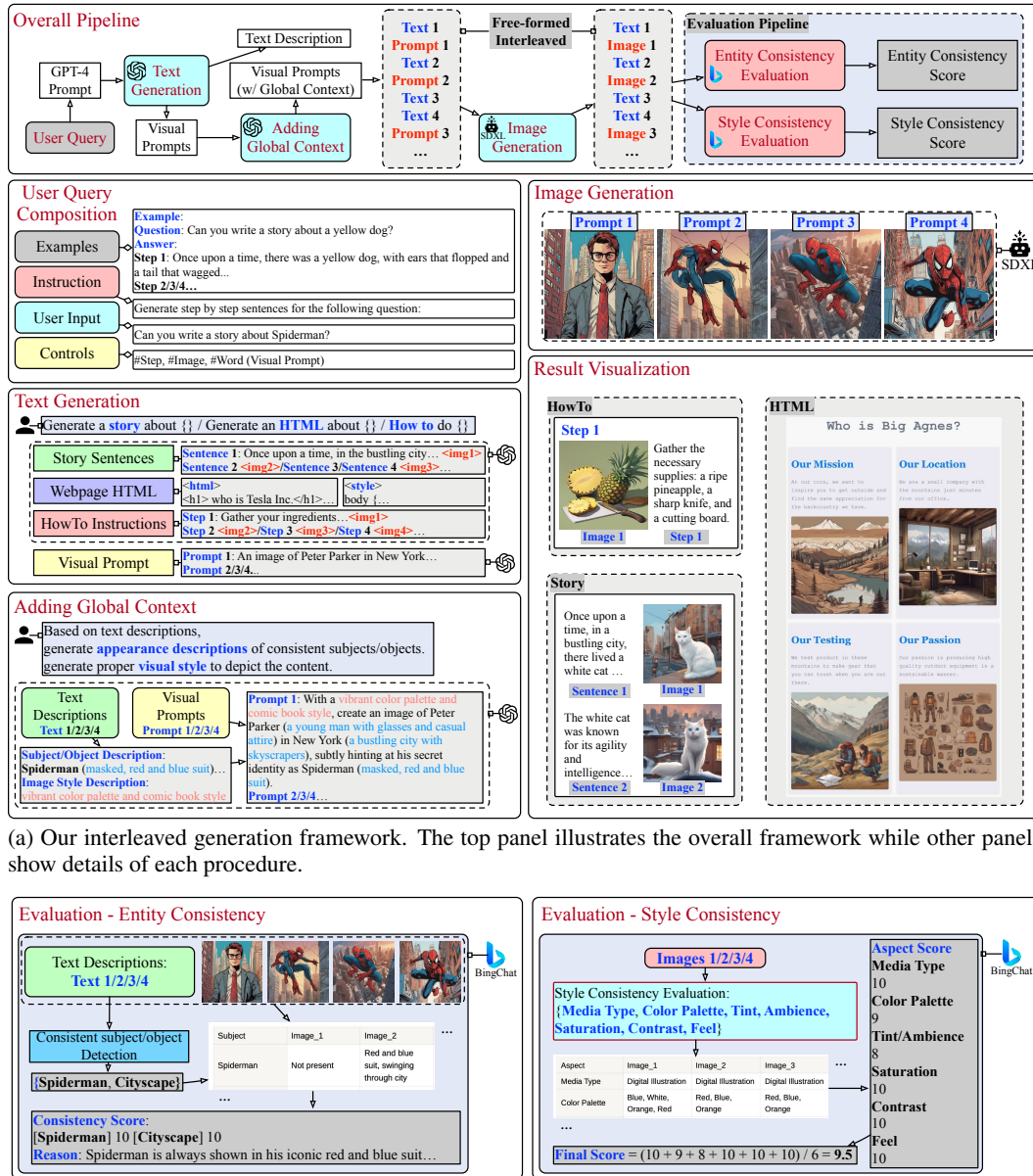
#### 3.1 INTERLEAVED CONTENT GENERATION

We achieve open-domain interleaved generation based on GPT-4 (OpenAI, 2023b) and SDXL (Podell et al., 2023). The top panel of Fig. 2a 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 into GPT-4, 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, SDXL converts visual prompts into real images, thereby creating the interleaved content.

**User Query Composition.** The input prompts to GPT-4 consist of four parts. We first add a few in-context examples to the start of the prompt. Each example shows the desired output corresponding to a specific input query. The in-context examples enable GPT-4 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-4 the desired output type, while the user input specifies the detailed content. Finally, we add a few control sentences to the prompt to control the number of image placeholders, story sentences, instruction steps, and `<div>`s in HTML, *etc.*

**Text Generation.** The first step of *OpenLEAF* is to generate text. By feeding the assembled prompt discussed in the previous part, we enable GPT-4 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. 2a, when generating stories and how-to instructions, GPT-4 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-4 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-4 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 SDXL to generate images. For the global entity context, we add a short appearance description of each common subject to the visual prompts, where GPT-4 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-4 to determine a proper visual style to depict the interleaved content, based on the generated text descriptions. For example, GPT-4 indicates that a vibrant color palette and comic book style are best 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



(a) Our interleaved generation framework. The top panel illustrates the overall framework while other panels show details of each procedure.

(b) Our LMM-based interleaved evaluation pipeline.











Figure 2: Overviews of the proposed interleaved generation framework: (a) and LMM-based evaluation pipeline (b).

generated by SDXL. Fig. 2a shows the process of adding global context into visual prompts. The visual prompts equipped with the global context are then converted into images by SDXL, resulting in the interleaved content.

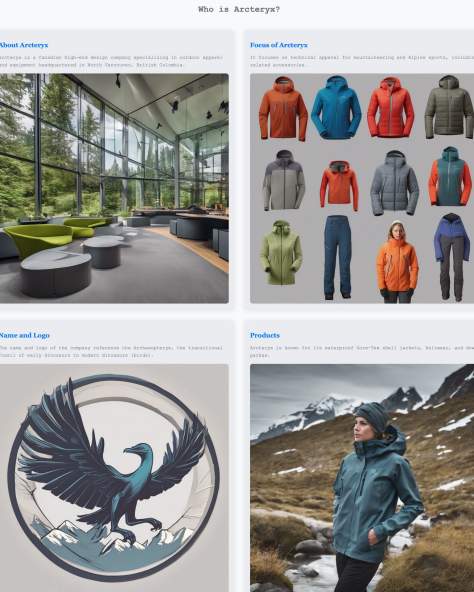
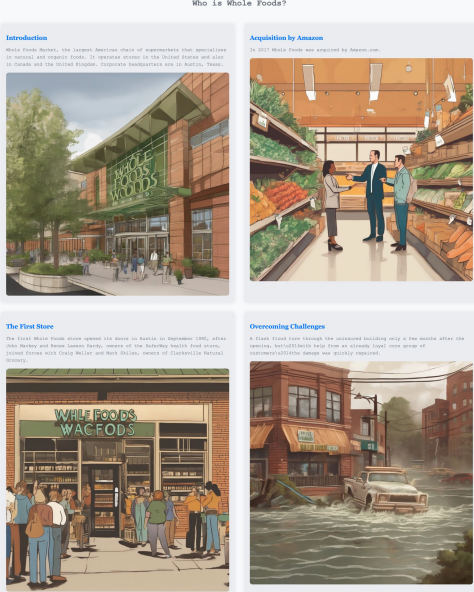
### 3.2 LMM-BASED EVALUATION.

We use BingChat (Microsoft, 2023) to evaluate the quality of interleaved content based on the chain-of-thought approach, with a focus on the entity and style consistencies.

**Entity Consistency Evaluation.** To evaluate the entity consistency of images within the interleaved content. We first prompt BingChat to detect two main common subjects from the generated text descriptions. For example, in Fig. 2b, the Spiderman and the background cityscape are the main subjects in images, which should have consistent entities and appearances between images. Subse-

<p><b>User Query:</b> Can you write a story about a <b>white cat</b>?</p>	<p><b>User Query:</b> Can you write a story about <b>the Hulk</b>?</p>
<p><b>Generated Interleaved Content:</b></p> <p>Once upon a ... as the sky.  The white cat was ... from above.  One day, the white cat ... alerting the firefighters.  Thanks to the white cat's ... city dwellers. </p>	<p><b>Generated Interleaved Content:</b></p> <p>Once upon ... the Hulk.  The Hulk was ... caused destruction.  One day, ... meteorite into pieces.  The city was saved, ... good of the city. </p>
<p><b>Story Visualization</b></p>  <ol style="list-style-type: none"> <li>Once upon a time, in a bustling city, there lived a white cat with fur as soft as snow and eyes as blue as the sky.</li> <li>The white cat was known for its agility and intelligence, often seen leaping from rooftop to rooftop, observing the city from above.</li> <li>One day, the white cat noticed a fire in one of the buildings. Without wasting a moment, it ran to the nearest fire station and started meowing loudly, alerting the firefighters.</li> <li>Thanks to the white cat's quick thinking, the firefighters were able to put out the fire and save the building. From that day onwards, the white cat was hailed as a hero and was loved even more by the city dwellers.</li> </ol>	<p><b>Story Visualization</b></p>  <ol style="list-style-type: none"> <li>Once upon a time, in a bustling city, there lived a man named Bruce Banner, who was also known as the Hulk.</li> <li>The Hulk was a giant green creature with immense strength and power, but he struggled to control his anger and often caused destruction.</li> <li>One day, a massive meteorite was heading toward the city. The Hulk, despite his uncontrollable anger, knew he had to save the city. He charged right into the sky and smashed the meteorite into pieces.</li> <li>The city was saved, and the Hulk was hailed as a hero. From that day on, he learned to control his anger and used his powers for the good of the city.</li> </ol>

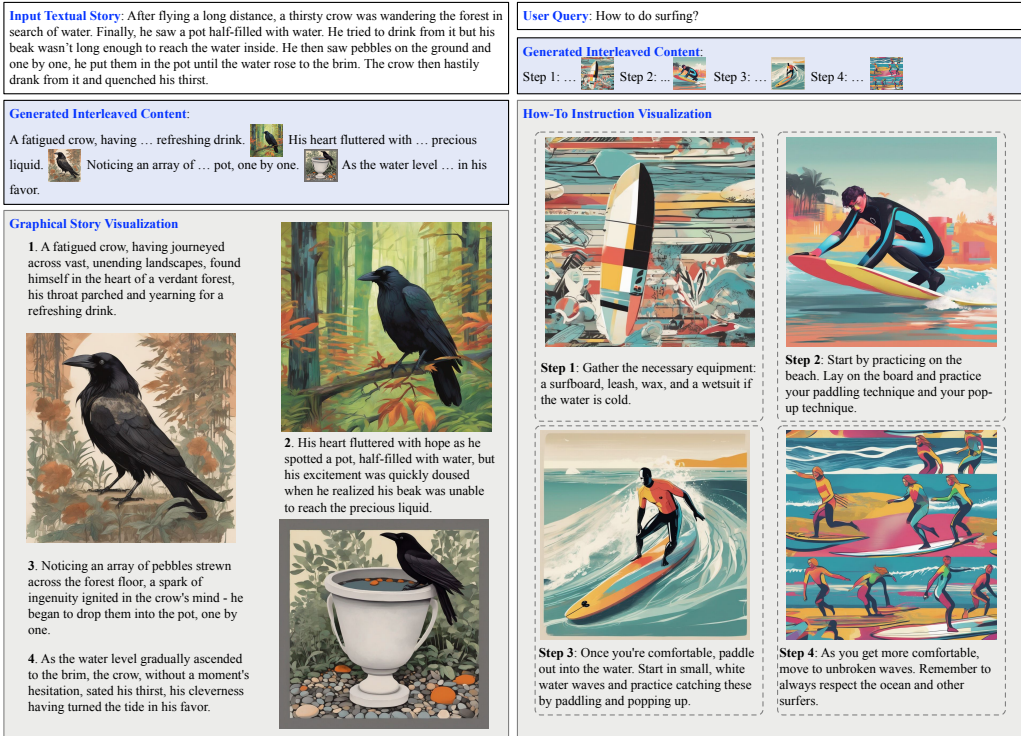
(a) Generating multi-modal stories.

<p><b>Input Textual Content:</b></p> <p><b>Title:</b> Who is Arcteryx?</p> <p><b>Content:</b> Arcteryx is a Canadian high-end design company specializing in outdoor apparel and equipment headquartered in North Vancouver, British Columbia. It focuses on technical apparel for mountaineering and Alpine sports, including related accessories. The name and logo of the company reference the Archaeopteryx, the transitional fossil of early dinosaurs to modern dinosaurs (birds). Arcteryx is known for its waterproof Gore-Tex shell jackets, knitwear, and down parkas.</p>	<p><b>Input Textual Content:</b></p> <p><b>Title:</b> Who is Whole Foods?</p> <p><b>Content:</b> Whole Foods Market, the largest American chain of supermarkets that specializes in natural and organic foods. It operates stores in the United States and also in Canada and the United Kingdom. Corporate headquarters are in Austin, Texas. In 2017 Whole Foods was acquired by Amazon.com. The first Whole Foods store opened its doors in Austin in September 1980, after John Mackey and Renee Lawson Hardy, owners of the Safeway health food store, joined forces with Craig Weller and Mark Skiles, owners of Clarksville Natural Grocery. Somewhat larger than a typical health food store, it offered a wider selection of food. A flash flood tore through the uninsured building only a few months after the opening, but with help from an already loyal core group of customers' the damage was quickly repaired.</p>
<p><b>Generated Webpage (From the generated HTML and images)</b></p> <p>Who is Arcteryx?</p> 	<p><b>Generated Webpage (From the generated HTML and images)</b></p> <p>Who is Whole Foods?</p> 

(b) Converting text introductions into graphical webpages.

quently, we input all images into BingChat, assigning a unique index  $image\_i$  to each image. We then have BingChat summarize the appearance of each common subject in the images and assign a score for the entity consistency based on the appearances of common subjects.

**Style Consistency Evaluation** We evaluate the style consistency of images based on seven visual factors: media type, color palette, tint, ambiance, saturation, contrast, and overall feel of images.



(c) Converting a text story into a multi-modal story.

(d) Generating visual how-to instructions.

Figure 3: Interleaved visual-language generation results of *OpenLEAF*. For story generation, story rewriting, and how-to instruction generation results, we show the raw interleaved content below the user query. Please zoom in on screen to see details.

The media type and feel recognize the global feeling of an image such as a realistic photo, painting, digital illustration, or cartoon. While the other evaluation factors detect more subtle visual style distinctions within images. Similar to the entity evaluation, we use BingChat to summarize the visual style of each image based on the pre-defined visual factors and score the style consistency for each factor. The final score is calculated as the mean value of the consistency score on all visual factors.

## 4 EXPERIMENTS

This section introduces our experimental settings, results, and analysis.

### 4.1 EXPERIMENT SETTINGS




**Implementation Details.** We utilize the text-only GPT-4 API of July 2023. For T2I generation, we use the text-to-image pipeline of the SDXL v1.0 model <sup>1</sup>, with all hyper-parameters set to their defaults. For LMM-Evaluation, we apply the precise mode of BingChat to both entity and style consistency evaluations. We develop a Python code to connect all the above prompting procedures, which automatically extracts the desired content from the previous model's output and forms the input for the next procedure.

**Compared Model Variants.** To demonstrate the effectiveness of the global context, we compare the performance of the proposed baseline and a simplified variant without the global context.

**Evaluation Set.** We collect a benchmark dataset of thirty problems, which covers four interleaved generation tasks: graphical storytelling, visual how-to instruction generation, text-to-graphic story

<sup>1</sup><https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0>

**Generated Interleaved Content:**

A fatigued crow, having ... refreshing drink.  His heart fluttered with ... precious liquid.  Noticing an array of ... pot, one by one  As the water level ... in his favor.

---

**Style Consistency Evaluation:**

**## User**  
please denote this image as image\_1.

**## Bing**  
Sure, we can refer to this image of the black crow perched on a branch in a forest as "image\_1". How can I assist you further with "image\_1"?

**## User**  
do these images have consistent visual styles between each other in terms of the media type, color palette, tint, ambience, saturation, contrast, and feel (painterly, digital, cartoony, or realistic)? Please give a table where each row is an aspect.

**## Bing**  
Sure, here is a comparison of the visual styles of the three images:

Aspect	Image_1	Image_2	Image_3
Media Type	Illustration	Digital Illustration	Illustration
Color Palette	Green, Orange, Black	Orange, Green, Brown, Gray	Beige, White, Black, Orange
Tint	Neutral	Warm	Neutral
Ambience	Peaceful and serene	Vintage and rustic	Peaceful
Saturation	Moderate	High	Moderate
Contrast	Moderate	High	Moderate
Feel	Semi-realistic	Semi-realistic	Graphic

From the table, it can be seen that all three images have a consistent visual style in terms of media type (illustration), ambience (peaceful), saturation (moderate), and contrast (moderate). However, they differ in terms of color palette, tint, and feel. Image\_1 and Image\_2 are semi-realistic while Image\_3 has a more graphic feel. The tint of Image\_2 is warm while the others are neutral. The color palettes also vary between the images.

**## User**  
In terms of the visual style consistency, can you give ONE single score for all images as a whole on each aspect, where 1 is the worst and 10 is the best. Please give a table where each row is an aspect. Please also show an equation about how the final score is calculated based on scores in the table.

**## Bing**  
Sure, here is a scoring of the visual style consistency of the three images:

Aspect	Score
Media Type	8
Color Palette	7
Tint	7
Ambience	9
Saturation	8
Contrast	8
Feel	7

The final score is calculated as the average of all the aspect scores. Here is the equation:  

$$SS\text{text}\{\text{Final Score}\} = \frac{\text{text}\{\text{Media Type Score}\} + \text{text}\{\text{Color Palette Score}\} + \text{text}\{\text{Tint Score}\} + \text{text}\{\text{Ambience Score}\} + \text{text}\{\text{Saturation Score}\} + \text{text}\{\text{Contrast Score}\} + \text{text}\{\text{Feel Score}\}}{7}$$
So, the final score would be:  

$$SS\text{text}\{\text{Final Score}\} = \frac{\text{text}\{8 + 7 + 7 + 9 + 8 + 8 + 7\}}{7} = 7.7155$$
So, on a scale of 1 to 10, the visual style consistency of these images scores a 7.71.

---

**Entity Consistency Evaluation:**

**## User**  
What are the consistent subjects shown in every step: "Step 1: A fatigued crow, having journeyed across vast, unending landscapes, found himself in the heart of a verdant forest, his throat parched and yearning for a refreshing drink."  
"Step 2: His heart fluttered with hope as he spotted a pot, half-filled with water, but his excitement was quickly doused when he realized his beak was unable to reach the precious liquid."  
"Step 3: Noticing an array of pebbles strewn across the forest floor, a spark of ingenuity ignited in the crow's mind - he began to drop them into the pot, one by one."  
"Step 4: As the water level gradually ascended to the brim, the crow, without a moment's hesitation, sated his thirst, his cleverness having turned the tide in his favor."

**## Bing**  
The consistent subjects in every step are:  
- The crow: The crow is the main character in all the steps. It is described as fatigued, hopeful, ingenious, and clever.  
- The forest: The forest is the setting for the story. It is described as verdant and has a pot and pebbles.  
- The pot: The pot is a key object in the story. It is half-filled with water and is initially out of the crow's reach.  
- The water: The water in the pot is what the crow desires. It is initially unreachable, but eventually drunk by the crow.  
- The pebbles: The pebbles are tools used by the crow to raise the water level in the pot. They are found on the forest floor and dropped into the pot one by one by the crow.

**## User**  
please denote this image as image\_1.

**## Bing**  
Sure, we can refer to this image as image\_1. How can I assist you further with image\_1?

**## User**  
does each subject have consistent appearances (shape, color) among all images that have the subject? Please give a table where each row is a subject. Please only list two main subjects.

**## Bing**  
Sure, here is a table comparing the appearances of the two main subjects, the crow and the forest, across the three images:

Subject	Image_1	Image_2	Image_3
Crow	The crow is black, facing left, and perched on a branch.	The crow is black, facing left, and perched on a branch.	The crow is black, facing left, and perched on an urn.
Forest	The forest is dense with trees and foliage in various shades of green and yellow.	The forest has trees and orange flowers in a muted color palette.	The background consists of green foliage.

From this table, we can see that both subjects have consistent appearances in terms of shape and color across all three images. The crow is always black and facing left, while the forest always includes green foliage.

**## User**  
In terms of the subject appearance consistency, can you give two scores for the first and second main subjects, respectively, where 1 is the worst and 10 is the best. Please take the shape and color consistencies into consideration. Please ignore images that do not contain a subject when giving scores. If a subject only appears in one image, please give 10.

**## Bing**  
Sure, here are the consistency scores for the two main subjects:

Subject	Consistency Score
Crow	10
Forest	8

The crow receives a score of 10 as its appearance is consistent across all images. It is always black and facing left.  
The forest receives a score of 8. While it consistently includes green foliage, the density and additional elements (like flowers) vary between images.

Figure 4: The entity and style consistency evaluation output of BingChat. BingChat can effectively generate rational responses, understand the meaning of pre-defined evaluation aspects, and give scores based on evidence.

rewriting, and webpage generation. The how-to instruction generation task has ten problems, which include the requests to generate the recipes for foods and the tutorials for sports and actions. The webpage generation task has five questions about converting a company introduction into a webpage or poster. The ten storytelling and five graphical story rewriting problems ask the model to generate animal- or superhero-related stories. Please refer to the supplementary material for problems of the constructed benchmark.

## 4.2 EXPERIMENT RESULTS

**Qualitative Results.** Fig. 3 shows the qualitative results of the *OpenLEAF*. We observe that *OpenLEAF* generates coherent interleaved image-text sequences based on arbitrary input queries. For example, in Fig. 3 (a), *OpenLEAF* generates coherent and interesting multi-modal stories about a white cat and the Hulk. All images have consistent entities and styles while being visually appealing. Fig. 3 (b) shows *OpenLEAF* generating commercial webpages based on introductions of “Arcteryx” and “Whole Foods”. *OpenLEAF* naturally splits text introductions into paragraphs and accurately adds paragraph headlines. The generated HTML and CSS files arrange the interleaved content in an intuitive and attractive way. Fig. 3 (c) shows *OpenLEAF* converting a textual story into multi-modal story and generating visual how-to instruction. The story sentences in the graphical story become more vivid compared with the input textual story while visually appealing and coherent images are naturally inserted into arbitrary text locations. The generated visual how-to instructions are informative and the illustration images are helpful in understanding terminologies of surfing.

More importantly, Fig. 4 shows that the BingChat is effective in evaluating entity and style consistencies, even for the complicated interleaved image-text sequences. For example, in evaluating entity consistency, BingChat accurately grounds common subjects extracted from text descriptions to the corresponding subjects in images. We observe that BingChat are capable of generating rationale text, understanding the meaning of and giving consistency scores on pre-defined sub-categories, as well as calculating the final score based on a clear mathematical equation.



Table 1: The mean and variance of the BingChat evaluation on the benchmark dataset. Adding global context improves the averaged consistencies and lowers the variances.

Model	Entity Consistency		Style Consistency	
	mean $\uparrow$	variance $\downarrow$	mean $\uparrow$	variance $\downarrow$
Ours w/o Global Context	7.84	1.22	8.00	0.88
Ours w/ Global Context	<b>8.40</b>	<b>0.77</b>	<b>8.22</b>	<b>0.79</b>

Table 2: The correlation analysis between the human and BingChat evaluations.  $p$  denotes the p-value.

Correlation Index	Entity Score	Style Score
Kendall’s Tau $\uparrow \in (-1, 1)$	0.87 ( $p = 0.0008$ )	0.58 ( $p = 0.0196$ )
Spearman’s Correlation $\uparrow \in (-1, 1)$	0.78 ( $p = 0.0080$ )	0.95 ( $p = 0.0000$ )

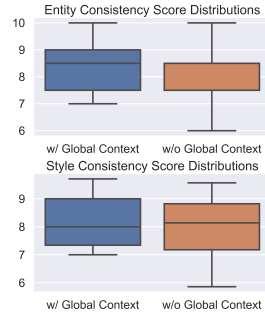


Figure 5: The distribution comparison of BingChat evaluation scores.

**Quantitative Results.** We then evaluate *OpenLEAF* and its variants using the effective BingChat evaluation. We compute the entity and style consistency scores on the interleaved content generated from problems in the benchmark set. Table 1 shows the mean and variance values of the entity and style consistency scores over the benchmark set while the boxplots in Fig. 5 show the comparison of the score distributions. Both *OpenLEAF* and its variant show good entity and style consistencies. After adding the global context, we observe that the mean values of both entity and style consistency scores increased, indicating improved consistencies. On the other hand, the global context also reduced the variances of both entity and style consistency scores, demonstrating that adding global context can improve the robustness of *OpenLEAF* in terms of content and style consistencies.

#### 4.3 EVALUATION PIPELINE ANALYSIS

To ensure that the LLM-evaluation pipeline can produce the entity and style consistency scores in line with the human perception. We conduct an analysis to study the score correlation between LLM-evaluation and the human rating. We collect ten image pairs from our generated interleaved content, where each pair comes from the same story. First, we use LLM-evaluation to give both entity and style consistency scores on each image pair. Next, we conduct a user study to let users rate the entity and style consistencies of each image pair. Considering that humans may have difficulty giving an absolute score of consistency, in the user study, we randomly select two image pairs from the evaluation set and show them side-by-side to users, letting users choose one pair that has better entity or style consistency. Finally, we calculate the consistency scores of each image pair based on the number of user preferences.

To determine whether the proposed LLM-evaluation aligns well with human perception, we compute Kendall’s Tau (Kendall, 1938) and Spearman’s Correlation (Kendall & Stuart, 1973) indexes, which measure the similarity between the rankings from the human and LLM-evaluation scores. Table 2 shows the correlation scores. For both entity and style consistencies, we achieve Kendall’s Tau and Spearman Rank-Order close to 1 with low p-values, strongly indicating that our LLM-evaluation method has a good alignment with the human evaluation in both measuring entity and style consistencies of the interleaved content.

## 5 CONCLUSION

In this paper, we fill the gap of open-domain interleaved image-text generation by introducing a baseline interleaved generation method *OpenLEAF* based on GPT-4 and SDXL, an evaluation pipeline based on BingChat, and a benchmark dataset to compare different approaches. Experimental results on the constructed benchmark dataset show that *OpenLEAF* has a strong ability to generate arbitrarily interleaved image-text sequences for addressing open-domain user queries. A comprehensive analysis based on the user study demonstrates that the evaluation method based on BingChat can effectively capture the styles and entities within images, thereby reliably evaluating generated interleaved multimodal contents.

**Ethics Statement.** In this research on the open-domain interleaved image-text generation, we uphold a commitment to ethical conduct guided by principles that prioritize human welfare, privacy, fairness, accountability, and transparency. Our model is built on top of GPT-4, BingChat, and Stable Diffusion XL, where all of the pre-trained models are publicly available and we believe their owners have a clear awareness of addressing the potential ethics issues. Our method does not bring new ethics issues. On the data side, we do not include any human-related problems into our constructed interleaved benchmark dataset to avoid the potential privacy and ethics issues.

**Reproducibility Statement.** All the pre-trained models we used are publicly available, where we clearly indicate the model version, settings, and hyper-parameters we used in our experiments. We believe our method and the evaluation pipeline are reproducible.

## REFERENCES

- Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.
- Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*, 2023.
- Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*, 2023.
- Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*, 2023.
- Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruva Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023.
- Google. Bard. <https://bard.google.com>, 2023. Accessed: 2023-07-17.
- Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14953–14962, 2023.
- G Kendall and A Stuart. The advanced theory of statistics: inference and relationship. griffin, 1973.
- Maurice G Kendall. A new measure of rank correlation. *Biometrika*, pp. 81–93, 1938.
- Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. Generating images with multimodal language models. *arXiv preprint arXiv:2305.17216*, 2023.
- Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M Rush, Douwe Kiela, et al. Obelisc: An open web-scale filtered dataset of interleaved image-text documents. *arXiv preprint arXiv:2306.16527*, 2023.
- Chunye Li, Liya Kong, and Zhiping Zhou. Improved-storygan for sequential images visualization. *Journal of Visual Communication and Image Representation*, pp. 102956, 2020.
- Chunyu Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. Multimodal foundation models: From specialists to general-purpose assistants. *arXiv preprint arXiv:2309.10020*, 2023.
- Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. Storygan: A sequential conditional gan for story visualization. In *CVPR*, 2019.
- Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. Visual-gptscore: Visio-linguistic reasoning with multimodal generative pre-training scores. *arXiv preprint arXiv:2306.01879*, 2023.

- Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, and Weidi Xie. Intelligent grimm–opened visual storytelling via latent diffusion models. *arXiv preprint arXiv:2306.00973*, 2023a.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023b.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. GpTeval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023c.
- Adyasha Maharana and Mohit Bansal. Integrating visuospatial, linguistic and commonsense structure into story visualization. *arXiv preprint arXiv:2110.10834*, 2021.
- Adyasha Maharana, Darryl Hannan, and Mohit Bansal. Improving generation and evaluation of visual stories via semantic consistency. *arXiv preprint arXiv:2105.10026*, 2021.
- Adyasha Maharana, Darryl Hannan, and Mohit Bansal. Storydall-e: Adapting pretrained text-to-image transformers for story continuation. In *ECCV*, 2022.
- Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.
- Microsoft. Bingchat. <https://www.microsoft.com/en-us/edge/features/bing-chat>, 2023. Accessed: 2023-09-25.
- OpenAI. Chatgpt, 2023a.
- OpenAI. Gpt-4 technical report, 2023b.
- Xichen Pan, Pengda Qin, Yuhong Li, Hui Xue, and Wenhui Chen. Synthesizing coherent story with auto-regressive latent diffusion models. *arXiv preprint arXiv:2211.10950*, 2022.
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.
- Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*, 2023.
- Yun-Zhu Song, Zhi Rui Tam, Hung-Jen Chen, Huihao-Han Lu, and Hong-Han Shuai. Character-preserving coherent story visualization. In *ECCV*, 2020.
- Quan Sun, Qiyang Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023.
- Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*, 2023.
- Gábor Szűcs and Modafar Al-Shouha. Modular storygan with background and theme awareness for story visualization. In *International Conference on Pattern Recognition and Artificial Intelligence*, pp. 275–286, 2022.
- Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.

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

Gangyan Zeng, Zhaohui Li, and Yuan Zhang. Pororogan: An improved story visualization model on pororo-sv dataset. In *Proceedings of the International Conference on Computer Science and Artificial Intelligence*, 2019.

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

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*, 2023.