

CoMM: A Coherent Interleaved Image-Text Dataset for Multimodal Understanding and Generation

Supplementary Material

Appendix

This supplementary document is organized as follows:

- Details of our dataset are shown in Sec. A.
- More experiment studies and analyses are shown in Sec. B.
- Ethical discussion and license are displayed in Sec. C.
- Filter strategy and evaluation details are in Sec. D.
- Model training details are shown in Sec. E.
- More visualization results are illustrated in Sec. F.

A. Details of Our Dataset

Data Source. As shown in Table 5, we collect our data from five sources. For StoryGen [23], we use only the original source images and employ Llama3 [25] to generate a more coherent and developed story text corpus. After filtering, CoMM contains 227K documents with 2.28M images and 139M text tokens. The ratio of different data sources is presented in Table 5. We randomly sample 500 documents each for the validation and test splits, with the remaining documents used for training.

Dataset Visualization. As displayed in Figure 4, we compared samples from the MMC4 [45] dataset and our CoMM dataset derived from various data sources. The comparison reveals: 1) MMC4 exhibits relatively poor style consistency. For the same “duck” entity, one image is in cartoon style while another is realistic (*cf.* Figure 4(a)). 2) Our CoMM dataset maintains high consistency in both entity representation and style across all data sources. 3) CoMM demonstrates enhanced narrative and stylistic diversity, covering different content orientations: Instructables focuses on narrative closure, StoryBird and StoryGen emphasize cartoon-style storytelling, while WikiHow and eHow concentrate on instructional content through illustrative methods.

Dataset Statistic. As shown in Figure 5, our dataset spans multiple domains, including technical fields (*e.g.*, technology, computing), creative pursuits (*e.g.*, crafting, painting techniques), and lifestyle areas (*e.g.*, healthy eating, personal experiences). The distribution of topics is relatively balanced, with no single category disproportionately dominating the dataset. These statistics demonstrate the diversity of our dataset.

B. More Experiment Studies and Analyses

Ablation study on the ITA. We conduct an ablation study on various thresholds for the image-text alignment (ITA). The results are shown in Table 7. We exclude data with an image-text alignment score below 4 in other experiments.

Ablation study on the ImgS. Table 8 displays the ablation study results of various image sequence (ImgS) scores. As seen, the score of 1.0 is best for filtering.

Ablation study on the MLLMs for dataset evaluation. To further verify the robustness of our dataset and evaluation results against model bias, we adopt Claude-3.5-Sonnet and human for the dataset quality evaluation process. The experimental results are shown in the Table 9. Since Claude-3.5-Sonnet is not used for dataset construction, the evaluation results are not affected by the bias of the model used for filtering. The differences in Claude-3.5’s results among the three datasets are consistent with GPT-4o’s results (*cf.* Table 1), where our dataset is superior to MMC4 and OBELICS in text development, completeness, and image-text alignment.

Scale law analysis. Table 10 shows our results trained with varying percentages of our dataset, following the same settings as Table 2. The evaluations are conducted on COCO and Flickr30K image captions in a 16-shot approach. Performance improves significantly as the data size of our dataset from 0% to 30%, highlighting the effectiveness of high-quality interleave data, and continues to improve gradually as data scale increases. Due to the high cost of acquiring high-quality interleaved image-text data, we cannot currently expand our dataset (with 2.28 million images) to billions. We believe that dataset quality is as important as scale.

Human study. To demonstrate the validity of the GPT-4o assessment, we further conduct a human study to assess the performance of the generation task in Table 11. Given time and labor constraints, we evaluate only the most challenging task, task 4 (Question-based Interleaved Image-Text Generation, *cf.* Sec.4.1). Results were collected from 32 people using the same criteria as in our paper. Humans are stricter and give lower scores than GPT-4o in evaluations, but the consistent score gap trend demonstrates the feasibility of using GPT-4o for evaluations.

Details and Additional Results Trained Using DPO [30]. We trained SEED-Llama [9] using the DPO [30] algorithm with a learning rate of 5×10^{-7} , a batch size of 16, and β in the DPO loss set to 0.05. The results for Task 1 are presented in our paper, and the results for the remaining tasks are shown in Table 6. These findings demonstrate that the current DPO algorithm can significantly enhance the performance of text generation tasks (*e.g.*, Task 1). However, its effectiveness in improving image generation quality remains limited, highlighting an area for further exploration and development in future research.

Ablation Studies among filter strategies. We conducted

| Data Source | Document Ratio | Image Ratio | Text Token Ratio |
|-----------------------|----------------|-------------|------------------|
| www.wikihow.com | 57.78% | 30.02% | 48.71% |
| www.ehow.com | 3.31% | 4.11% | 2.53% |
| storybird.com | 1.54% | 1.22% | 0.57% |
| StoryGen [23] | 2.19% | 1.76% | 1.90% |
| www.instructables.com | 35.18% | 62.89% | 46.28% |

Table 5. Collected data source of CoMM.

(c) Sample from MMC4

Do you remember the Fairy Story of the Ugly Duckling?



The over-sized, clumsy chick who came out of a duck's nest and grew up thinking there was something wrong with him because he didn't look or act like his siblings and got made fun of.I then found my fellow "Swans" when I began studying Counselling and then NLP coaching where I learned to celebrate my differences as my unique gifts to the world.

I now realise that the reason I felt different to others was because I WAS different (in a good way) from the relatively small group of people I came across in School and work. Not that there was anything wrong with them, but they had different skills and attitudes to life.Please use the free resources on my web-site, Instagram, Twitter and Facebook pages to help you continue your journey, and sign up for my free on-line course "10 Steps to Calm" for more in-depth information about my journey.

(b) Sample from CoMM (Instructables)

Quintessential Way to Knot Your Thread When Sewing Simple thing, right? Thread your needle and then put a knot at the end.....So, here's an Instructable about a trick that saved my sanity and opened up a new world (I could have never started to do Hawaiian quilting, which is all handsewn, without it) of enjoyment.

Step 1. First, Thread your needle. Now!: Hold an end of the thread between your index finger and the needle, near the sharp end.

Step 2. Wrap the thread around the needle three times.

Step 3. Grip the wrapped thread between the thumb and index finger of the other hand.

Step 4. Let the needle pass your fingertips while maintain a grip on the wrapped thread region. Pull the thread until the wrapped region get to the end of the thread. And so! You have a knot perfectly placed at the end of your thread and your hands aren't all cramped up, or anything. Easy-peasy!







(c) Sample from CoMM (StoryBird)

Guide me over mountain tops and through the vast unknown, Beneath your wings of grandeur as you gather me back home.



For I was lost and wandering parched from all the dust, battling to quench my thirst for this I knew I must.



Days and days I staggered round in circles all alone, Hoping you would find me for my flesh was near to bone.



But just as I was starting to lose hope of being found, You lifted me to safety without a single sound.



And all at once I felt such love as I had never known, Beneath your wings of grandeur as you gathered me back home.



(d) Sample from CoMM (StoryGen)

In the world of imagination, a magical book was born. The cover featured a bright blue rectangle with a vibrant, multi-colored star, and a serene green tree with a sturdy brown trunk. The author and illustrator, Timmy Yan, had poured his heart into creating this enchanting tale.

Anya and Tommy, two curious friends, stumbled upon the magical book on a sunny day. As they stood next to each other, the book began to radiate a warm, blue light, filling them with wonder and excitement. The trees nearby seemed to lean in, as if to listen to the whispers of the book.

In the park, Anya and Tommy discovered the magic book, and their lives were forever changed. They met a trio of fantastical creatures: a tall, symbol-adorned being; a round-headed, gentle soul; and a blue, square-faced guardian. The air was electric with anticipation as the friends prepared to embark on an unforgettable adventure.

With the magic book open, Anya and Tommy found themselves transported to a new world. They stood side by side, beaming with joy, as a brown box with a colorful star sat beside them. The box seemed to pulse with an otherworldly energy, guiding them deeper into the heart of the book.

As they journeyed through the book, Anya and Tommy encountered a benevolent monster, who gifted them a powerful, shining weapon. With the monster's guidance, they traversed the pages, following the path to the book's final destination. The star above them shone bright, illuminating their path to "The END."







(e) Sample from CoMM (WikiHow)

Prepare a cup of warm water. You'll want a nice wide bowl.



Add some liquid soap. Stir it together, and carefully place your rock inside the mix. But before you do, check the rock's type. Some rocks may not be able to go in water.



Use a sponge or a cloth to scrub your pet rock inside with the soap-water mixture. If you can't use any of these, use your hands. Scrub for 2-4 minutes. Make sure it rubs all over, under any crevices.



Dry your pet rock with a small cloth. Now, let your pet rock rest. Bathing is very important when "pet rocking", and rocks like to contemplate their day after a relaxing bath. You may need to redo any text/pen that you may have drawn on.



(f) Sample from CoMM (eHow)

Step1. Work out a mosaic design. Roman mosaics depicted gods and goddesses, gladiators, birds and a wide range of other designs. Parents can tell children a few Roman myths and allow the children to draw the design in pencil on the black paper based on the myth. For example, parents might tell the story of Cupid and his wife Psyche or they might tell stories of a gladiator fighting a lion. The Holiday Spot (see Resources) has a brief summary of Cupid's story available. Let the children work on the drawing based off the story.

Step2. Cut the paper into small pieces. Education.com suggests about 1/4 inch to one inch square pieces, depending on the child's preferences and age. Use appropriate child safety scissors for any children cutting the paper. Cut strips and then cut the strips down into squares.

Step3. Glue the pieces of paper to the black construction paper. Allow the children to use a glue stick and put the paper on the background according to his or her preferred color combinations. Education.com suggests starting from the outside, outlining the drawing and then working inward.

Step4. Allow the glue to dry and the mosaic is complete.






Figure 4. Comparison of samples from different datasets. (a) from the MMC4 [45] dataset; (b)-(f) from different data sources within CoMM (Ours) dataset.

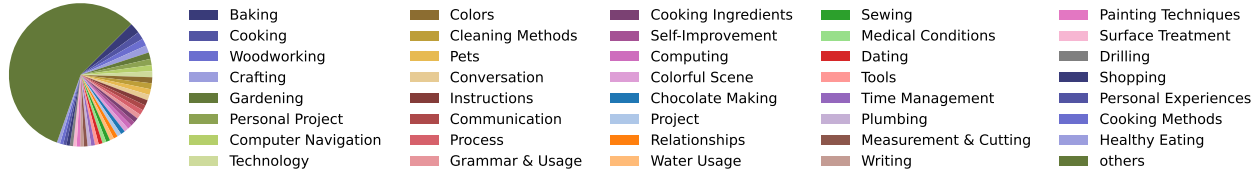


Figure 5. Topic visualization of our dataset. ‘Others’ contain ‘Exercise’, ‘Drawing & Design’, ‘Boating’, etc., totaling 144 topics.

| Methods | Size | T2I Sequence Generation (Task2) | | | | |
|----------------|------|---------------------------------|-------|-------|------|------|
| | | FID ↓ | IS | SSIM | PSNR | IRS |
| SEED-Llama [9] | 8B | 57.96 | 8.40 | 20.53 | 7.87 | 5.27 |
| + DPO [30] | 8B | 63.05 | 8.28 | 20.99 | 8.05 | 4.69 |
| SEED-Llama [9] | 14B | 66.23 | 11.06 | 20.83 | 8.12 | 6.24 |
| + DPO [30] | 14B | 72.60 | 9.98 | 20.49 | 7.84 | 5.75 |

| Methods | Size | Continuation Generation (Task3) | | | | | |
|----------------|------|---------------------------------|--------|-------|------|------|------|
| | | Style | Entity | Trend | CPL | ImgQ | IRS |
| SEED-Llama [9] | 8B | 6.28 | 5.84 | 5.72 | 6.28 | 6.55 | 2.92 |
| + DPO [30] | 8B | 6.19 | 5.58 | 5.23 | 5.74 | 6.10 | 2.86 |
| SEED-Llama [9] | 14B | 6.68 | 6.22 | 6.13 | 6.66 | 6.67 | 3.23 |
| + DPO [30] | 14B | 6.05 | 5.54 | 5.26 | 5.97 | 6.09 | 3.11 |

| Methods | Size | Question-based Generation (Task4) | | | | | |
|----------------|------|-----------------------------------|--------|-------|------|------|------|
| | | Style | Entity | Trend | CPL | ImgQ | IRS |
| SEED-Llama [9] | 8B | 7.55 | 6.81 | 6.15 | 5.13 | 6.36 | 1.46 |
| + DPO [30] | 8B | 7.94 | 6.63 | 5.25 | 4.61 | 5.95 | 1.69 |
| SEED-Llama [9] | 14B | 7.51 | 6.61 | 6.30 | 6.13 | 6.66 | 2.50 |
| + DPO [30] | 14B | 7.54 | 6.44 | 5.40 | 5.47 | 6.13 | 2.76 |

Table 6. Performance results of SEED-Llama [9] trained by DPO [30] in our preference dataset. ‘CPL.’ stands for Completeness, ‘ImgQ’ indicates Image Quality, ‘IRS’ means Illustration Relevance Score, and ‘↓’ denotes that lower values are better.

| ITA score | 2 | 3 | 4 | 5 |
|-----------|-------|-------|-------|-------|
| COCO | 103.1 | 107.1 | 109.3 | 109.1 |
| Flickr30K | 69.3 | 70.7 | 71.5 | 70.9 |

Table 7. Ablation study on various image-text alignment (ITA) thresholds for data filtering. We train the model with filtered data and evaluate it using a 16-shot image caption way.

ablation studies on our dataset. The results show that our high-quality data boosts performance, and our filtering strategies further enhance performance. ‘Origin’ refers to using our collected data without filtering. ‘ITA’ involves using LLM combined with a caption model to filter out text sequences and poorly aligned image-text data, while ‘ImgS’ uses CLIP to filter image sequences. Consistent with the few-shot experiment in Table 12, we used 16 shots.

C. Ethical Discussion and License

Ethical Discussion. Collecting data from online sources comes with the risk of encountering content that may not be suitable for all audiences. Fortunately, this risk is minimized

| ImgS Score | 0.5 | 1.0 | 1.5 | 2 |
|------------|-------|-------|-------|-------|
| COCO | 108.5 | 109.3 | 108.7 | 109.2 |
| Flickr30K | 70.2 | 71.5 | 71.6 | 71.3 |

Table 8. Ablation study on various our proposed image sequence (ImgS) score thresholds for data filtering. We train the model with filtered data and evaluate it using a 16-shot image caption way.

| Models | DLP | CPL | ITA |
|-------------|--------------------|--------------------|--------------------|
| | CL / HM | CL / HM | CL / HM |
| MMC4 | 4.61 / 6.20 | 4.76 / 6.48 | 4.30 / 5.79 |
| OBELICS | 5.35 / 5.58 | 5.28 / 5.60 | 3.22 / 4.87 |
| CoMM (Ours) | 7.58 / 8.32 | 7.12 / 8.57 | 8.29 / 8.56 |

Table 9. Dataset quality evaluated by Claude-3.5-Sonnet (CL) and Human (HM). ‘DLP’ stands for Development, ‘CPL’ signifies Completeness, and ‘ITA’ represents Image-Text Alignment. For CL, Each dataset is randomly sampled with 5000 cases for evaluation. For HM, a total of 570 feedback responses were collected from 19 persons, each evaluating 30 documents.

in our case because we focus on high-quality data, such as instructional steps and visual stories. Besides, these specific websites have their review/ editorial processes, which significantly improve data quality and reduce potential hazards. For example, WikiHow [42] claims that ‘the average WikiHow article has been edited by 23 people and reviewed by 16 people’. To further ensure the integrity of our dataset, we perform a rigorous screening process to filter out any NSFW content (as mentioned in Sec 3.1), trying to maintain a clean, and reliable dataset suitable for all users.

License and Author Statement. We release this dataset under a CC-BY license and Terms of Use that require disclosure when used for model training. This license does not override the original content licenses; all use must comply with the original licenses and data subjects’ rights. We clarify the user’s responsibilities and liabilities here. While we’ve tried our best to ensure data accuracy and legality, we cannot guarantee absolute correctness. We assume no liability for rights violations, including but not limited to copyright, privacy issues, or misuse of sensitive information.

By using this dataset, you accept full responsibility for legal or other consequences. You agree to adhere to all relevant laws, regulations, and ethical guidelines. Accessing or using this dataset signifies your acceptance of this statement and

| Percent (Image Number) | 0% (Baseline) | 30% (0.68 M) | 60% (1.37 M) | 100% (2.28 M) |
|------------------------|---------------|--------------|--------------|---------------|
| COCO | 98.8 | 105.3 | 106.9 | 109.3 |
| Flickr30K | 62.8 | 68.7 | 70.8 | 71.5 |

Table 10. Scale law analysis.

| Models | Size | Style | Entity | Trend | CPL | ImgQ |
|----------------|------|-------|--------|-------|------|------|
| MiniGPT-5 [44] | 7B | 6.39 | 5.92 | 5.84 | 6.67 | 6.16 |
| SEED-Llama [9] | 8B | 6.79 | 6.48 | 5.48 | 6.68 | 6.24 |
| SEED-Llama [9] | 14B | 6.83 | 6.37 | 5.54 | 6.37 | 6.33 |
| Emu2 [36] | 33B | 7.27 | 6.56 | 5.47 | 6.79 | 6.03 |

Table 11. Human study on question-based interleaved image-text generation task.

| Origin | ITA | ImgS | COCO | Flickr30k | TextVQA |
|--------|-----|------|-------|-----------|---------|
| | | | 98.8 | 62.8 | 27.3 |
| ✓ | | | 105.7 | 67.4 | 31.5 |
| ✓ | ✓ | | 108.5 | 70.2 | 34.2 |
| ✓ | | ✓ | 107.7 | 68.3 | 33.6 |
| ✓ | ✓ | ✓ | 109.3 | 71.5 | 35.8 |

Table 12. Ablation studies among filter strategies.

| Dataset | Style | Entity | Trend | CPL | ImgQ | IRS |
|------------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Continuation Generation (Task 3) | | | | | | |
| MMC4 | 5.22 | 4.90 | 4.47 | 4.45 | 5.5 | 1.32 |
| OBELICS | 4.67 | 4.10 | 3.83 | 3.58 | 5.26 | 0.96 |
| CoMM | 6.68 | 6.22 | 6.13 | 6.66 | 6.67 | 3.23 |
| Question-based Generation (Task 4) | | | | | | |
| MMC4 | 5.78 | 4.71 | 3.58 | 3.04 | 4.31 | 1.31 |
| OBELICS | 3.25 | 2.65 | 1.82 | 1.84 | 4.59 | 1.14 |
| CoMM | 7.51 | 6.61 | 6.30 | 6.13 | 6.66 | 2.50 |

Table 13. Performance comparison among three datasets on Task 3 & 4. We train different datasets on SEED-Llama-14B. We used LLM to generate titles for MMC4 and OBELICS, which served as pseudo-label questions for training.

the CC-BY license terms. Disagreement with these terms means you are not authorized to use the dataset.

D. Filter Strategy and Evaluation Details

D.1. Data Quality Filter Prompt.

Below is the prompt for ensuring data quality in text and image-text alignment. When using GPT-4o [27], which can see images directly, we input the original images directly. However, when using Llama3 [25], which cannot see images, we first employ CogVLM [40] to convert the image into a detailed caption, then input it in the “<IMAGE>image description</IMAGE>” format.

You are a master of multi-modal evaluation.
Your task is to evaluate the quality of
a docs that contains images and text.

Images will be presented in the format <IMAGE>image description</IMAGE>. The textual content will be presented as plain text. Evaluate the following criteria:

1. **Development:** Assess the coherence and logical flow of the data. Only the most logically consistent and well-integrated contexts should receive high scores.
2. **Completeness:** Check if the content provides a comprehensive and detailed overview of the topic. Full scores should only be given for thorough and exhaustive coverage.
3. **Interleaving of Images and Text:** Ensure that the images and text are perfectly aligned. Discrepancies or inconsistencies should result in significant deductions.

Scores should range widely to highlight exceptional quality or notable deficiencies. Each criterion should be evaluated and concluded with a score on a scale from 0 to 10, where 0-2 indicates major deficiencies and 8-10 indicates exemplary performance. Structure your response as follows:

```
<Development>
  <Problem>Brief description of any issues
  </Problem>
  <Score>Numerical rating</Score>
</Development>
<Completeness>
  <Problem>Brief description of any gaps</
  Problem>
  <Score>Numerical rating</Score>
</Completeness>
<Image-Text Interleaving>
  <Problem>Brief description of any
  discrepancies</Problem>
  <Score>Numerical rating</Score>
</Image-Text Interleaving>
```

Emphasize the identification of particularly strong or weak points in the Problem section. This feedback will guide you to adjust scores to be more polarized, reflecting a clear distinction between high and low quality .

```
Data to Review:
<data>
{}
</data>
```

D.2. Evaluation Prompt for Interleaved Generation Content.

We explain the motivation and detailed prompt design of the GPT-4o evaluation here.

Document Completeness, Image Sequence Coherence and Image Quality. Here are the prompts for evaluating the document completeness, image sequence coherence, and image quality.

We are evaluating the results of a model designed for generating interleaved image-text documents. The model's input, starting with "INPUT:", can either be the beginning of a text-image interleaved document or a specified topic. Its output, starting with "OUTPUT:", will then be either a continuation of the document or content generated based on the given topic. The image with the index i will be enclosed by the symbols "<Img_i>" and "</Img_i>". The images are numbered sequentially from 0 to N (including the input images).

As an expert in multimodal evaluation, your task is to assess the quality of the output that includes both images and text. The images are numbered sequentially from 1 to n (include the input images). Use the guidelines below to assign a final score.

Scoring Guidelines:

- 0-3: Major deficiencies, misalignment, or inconsistency
- 4-7: Minor gaps, misalignment, or inconsistency
- 8-10: Complete and thorough alignment, strong consistency

Scoring Criteria:

1. Image Coherence:

- Evaluate the consistency of style and entity between the output images. Assess whether the trend shown by the image sequence aligns with the text. Finally, an overall consistency score will be assigned to the image sequence.

2. Completeness:

- Summarize the output document's topic and evaluate how thorough and comprehensive the output content is. Evaluate Thoroughness and Comprehensiveness:
 - Is the text content complete? Is there anything missing?
 - Is the image content complete? Are any images missing?
 - Do the images and text fully support each other? Is there any missing image or text?

3. Image Quality:

- Evaluate the quality of the output images based on the following aspects:
 - Realism: Determine whether the image resembles a real scene or object and identify any signs of artificial model synthesis.
 - Completeness: Check if the objects in the image are fully intact, without any noticeable missing parts, truncation, or damage.
 - Clarity: Determine if the details are sufficient and if the image is free of blurriness or out-of-focus areas.
 - Composition balance: Evaluate the aesthetic quality and balance of the image composition, ensuring that the main subjects are well-framed and the composition is visually pleasing.

Assume the index of the first image in the output is K.

JSON Output Structure:

```
{
  "scores": {
    "Image_Coherence": {
      "pair_scores": {
        "image_K_and_K+1": {
          "style_consistency":
            0-10,
          "entity_consistency":
            0-10,
          "justification": "Brief
            explanation of any
            gap"
        },
        "image_K+1_and_K+2": {
          "style_consistency":
            0-10,
          "entity_consistency":
            0-10,
```



```

        "justification": "Brief
            explanation of any
            gap"
    },
    // Continue for remaining
    pairs...
},
"overall_score": {
    "style_consistency": 0-10,
    "entity_consistency": 0-10,
    "trend_consistency": 0-10,
    "overall_consistency":
        0-10,
    "justification": "Brief
        explanation of overall
        consistency"
},
},
"Completeness": {
    "Summarize": "brief summary",
    "Justification": "brief
        justification of any issue",
    "Score": 0-10
},
"Image_Quality": {
    "Score": 0-10,
    "Justification": "brief
        justification of any
        deficiencies in image
        quality",
}
}
}

```

Data to Review:

Illustration Relevance Score. In our evaluation tests, we found that GPT-4o has a good understanding of images, but the text context will seriously influence this understanding. For example, when only images from a step-by-step instruction document are input, GPT-4o can accurately describe both the correctly ordered and reversed-ordered image contents. However, when the interleaved texts and images are input together, GPT-4o tends to produce similar descriptions for both the correctly ordered and reversed-ordered images, which is incorrect.

Thus we design an evaluation process similar to human document creation to mitigate this limitation. We first generate the required image content based on the text context and then evaluate the consistency between the image descriptions and the corresponding images. Specifically, this process involves two model invocations, the following are the prompts for the first invocation:

We are evaluating the results of a model designed for generating interleaved

image-text documents. The model's input, starting with "INPUT:", can either be the beginning of a text-image interleaved document or a specified topic. Its output, starting with "OUTPUT:", will then be either a continuation of the document or content generated based on the given topic. The image with the index i will be enclosed by the symbols "<Img_i>" and "</Img_i>". The images are numbered sequentially from 0 to N (include the input images). Now we hide the output's images while preserving the "<Img_i></Img_i>". As an expert in multimodal evaluation, you are responsible for predicting the removed image's content based on the input and the output text context.

Tasks:

1. Predict Each Image's Content:

For each image content prediction, predict the most probable and suitable image content based on the input and text context in the output. The description should consider the illustration needs (What should the image illustrate to complement its surrounding text context?), content description (Provide a detailed description of what the image should contain.), and context coherence (Ensure that the final narrative flows well and forms a complete, coherent document.).

Assume the index of the first removed image in the output is K.

JSON Output Structure:

```

{
  "Tasks": {
    "Create an Interleaved Text-Image
    Document": {
      "Content of Image K": "
        predicted content of image K
      ",
      "Content of Image K+1": "
        predicted content of image K
        +1",
      ...
      "Content of Image N": "
        predicted content of image N
      ",
    }
  }
}

```

Data to Review:

After the first invocation, we reorganize the output descriptions with corresponding images and start the second model invocation with the following prompts:

```
As an expert in image description
evaluation, your job is to assess the
consistency between two sets of images
and their corresponding descriptions.
Use the criteria below to assign a final
score.

The input will be formatted as description-
image pairs like <Description_i> image
description </Description_i> <Img_i>
image </Img_i>. Note that sometimes one
of the descriptions and the image is
missing, just score that input data as
0!

Scoring Guidelines:
0-3: Major deficiencies/misalignment/
inconsistency,
4-7: Minor gaps/misalignment/
inconsistency,
8-10: Complete and thorough alignment,
strong consistency.

Scoring Criteria:

1. Consistency:
- Task: Evaluate the consistency
between each image and its
corresponding description.

JSON Output Structure:
{
  "Consistency": {
    "image_1_score": 0-10,
    "image_2_score": 0-10,
    ...
    "image_n_score": 0-10,
    "overall_score": 0-10,
    "Justification": "Brief justification
of any issue identified"
  }
}

Data to Review:
```

The output overall score is the IRS.

E. Model Training Detail

MiniGPT-5 [44] combines the Stable Diffusion with LLMs through “generative vokens”. This model adopts a two-stage training strategy tailored for description-free multimodal generation. Initially, it focuses on extracting high-quality text-

aligned visual features. In the subsequent stage, it ensures optimal coordination between visual and textual prompts, significantly enhancing its ability to generate coherent multi-modal content.

Training Settings. We train MiniGPT-5 using 8 A100-80G GPUs, fine-tuning the parameters of the LoRA [14] layers (the rank is 32) in the LLM backbone and the Feature Mapper for output visual tokens. The learning rate is set to $5e-5$ for the LoRA layers and $5e-4$ for the other trainable parameters, with a total of 5 training epochs. All other settings follow those of MiniGPT-5.

SEED-Llama [9] equips the pre-trained LLM [39] with a VQ-based image tokenizer (SEED), which processes images into discrete tokens. This tokenizer utilizes a 1D causal dependency to align visual tokens with the autoregressive nature of LLMs, enhancing semantic coherence between text and images. Enhanced by extensive multimodal pretraining and fine-tuning under a next-word-prediction objective, SEED-Llama excels in handling both comprehension and generation tasks within a unified multimodal framework.

Training Settings. We train SEED-Llama-8B and SEED-Llama-14B using 8 A100-80G GPUs. Only the parameters of the LoRA [14] layers (with a rank of 16) in the LLM backbone are fine-tuned. The learning rate is set to $1e-5$, and the training consists of 10,000 steps.

Emu2 [36] is a generative multimodal model, trained on large-scale multimodal sequences with a unified autoregressive objective. This model showcases significant capabilities in multimodal in-context learning, adept at complex tasks that require on-the-fly reasoning, such as visual prompting and object-grounded generation.

Training Settings. Emu2 is trained using 16 A100-80G GPUs. We fine-tune the parameters in the linear projection layer for input and output visual embeddings, as well as the LoRA [14] layers (with a rank of 32) within the LLM backbone. The learning rate is set to $5e-5$, and the training lasts for 5 epochs.

F. More Generation Visualization

Qualitative Analysis of Interleaved Generation We visualized the results of three baseline models (Emu2 [36], SEED-Llama [9], and MiniGPT-5 [44]) across four interleaved generation tasks: image-to-text sequence generation (*cf.* Figure 6), text-to-image sequence generation (*cf.* Figure 7), interleaved image-text content continuation (*cf.* Figure 8), and question-based interleaved image-text Generation (*cf.* Figure 9). From the results, we can observe that: 1) For the single textual modality generation, the Emu2 model can more accurately describe entities (*e.g.*, tripod, straw, and sunglasses in Figure 6) appearing in images, producing more coherent and concise text descriptions. 2) For the single visual modality generation, SEED-Llama can generate images that are highly aligned with texts and exhibit consistent style,

as exemplified by the uniformity in shape and decoration of the pie depicted in Figure 7. 3) For interleaved image-text generation, SEED-Llama also exhibits more coherent and stylistically consistent image-text outputs. Conversely, the Emu2 model generates images with excessive uniformity, which detracts from their developmental progression. While MiniGPT displays some developmental aspects, it suffers from inconsistencies in style.

In addition, to demonstrate that MLLMs can facilitate a variety of applications with the interleaved multimodal generation ability, we visualize the SEED-Llama’s generation results of visual storytelling with MMC4 and our CoMM, respectively (*cf.* Figure 10). As seen, compared to MMC4, our CoMM enables the SEED-Llama to generate stories with more consistent style and coherent content.

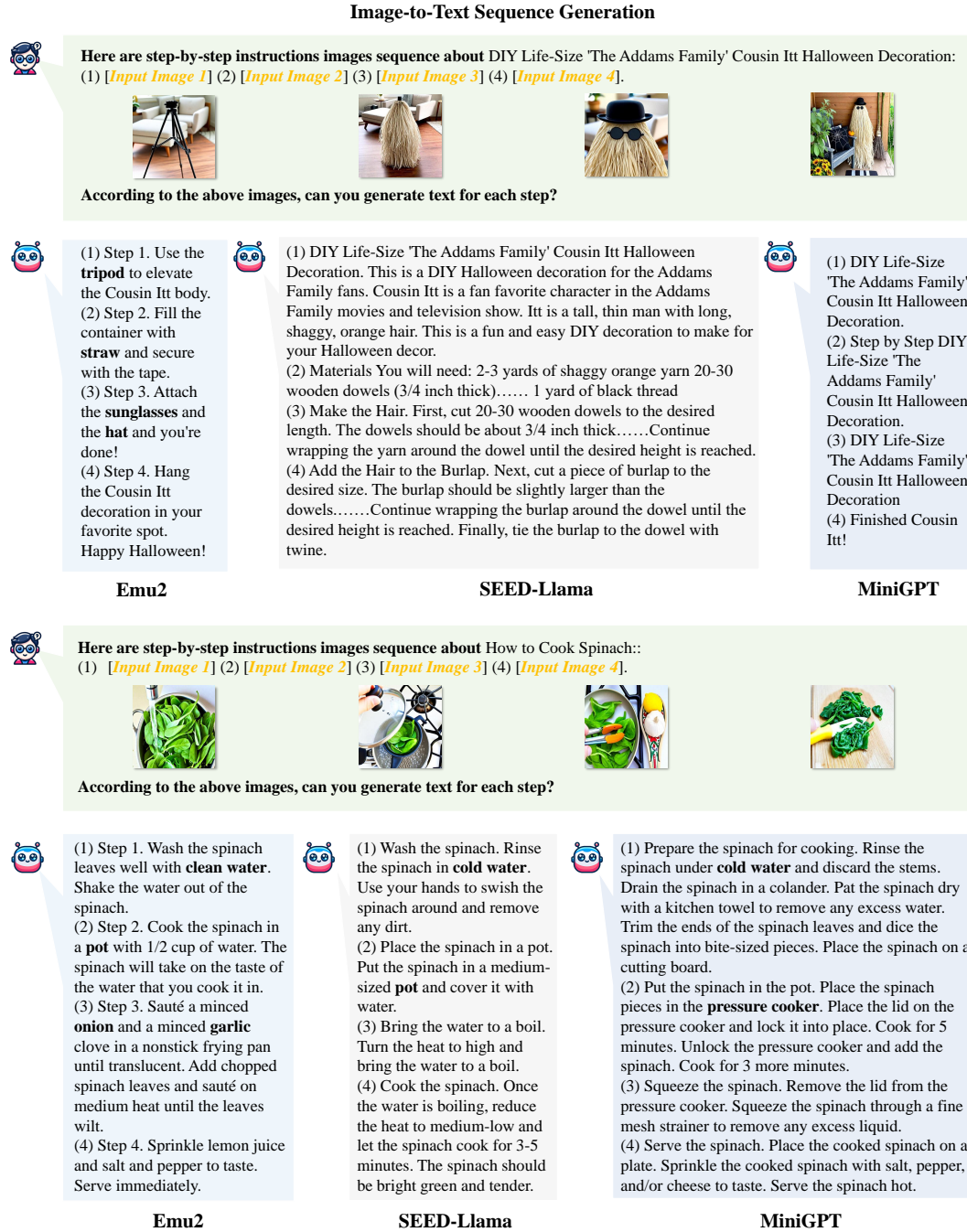


Figure 6. Visualization of image-to-text sequence generation from Emu2 [36], SEED-Llama [9], and MiniGPT-5 [44], separately.

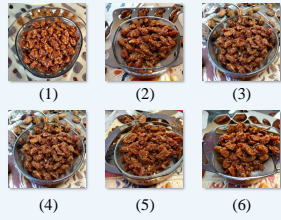
Text-to-Image Sequence Generation



Here are step-by-step instructions without images about Spicy Wild Cranberry Chutney:

- (1) Spicy Wild Cranberry Chutney. Low-bush wild cranberries (also called lingonberries) are my favourite berry. They don't have a pit like the high-bush cranberry so they are easy to use for many different recipes. This is my take on a classic cranberry chutney - adding jalapeno and lime to kick it up a notch. This recipe makes three 1-cup mason jars of chutney and takes about an hour - prep to finish (not including the time spent berry picking).
- (2) Berries. Pick the wild cranberries and remove any leaves and tag-alongs. It works well to float the berries in water as the impurities will sink. It also works well to lay the berries out on a terry towel and roll them into a bowl. The leaves and other things will stick to the towel. Wash the berries well.
- (3) Ingredients 8 shallots, coarsely chopped 1 tablespoon oil (I used coconut) 4 cups fresh cranberries 1/2 cup brown sugar, loosely packed 1/4 cup cider vinegar 6 cloves minced garlic 1 inch peeled fresh ginger 1 teaspoon salt 1 teaspoon black pepper 1 fresh jalapeno (seeded, unless you like more spice) Juice and zest of 2 limes
- (4) Cook the Chutney. Cook the shallots first in the coconut oil over moderate heat, stirring occasionally, until softened. Using a food processor, blend the garlic, ginger and jalapeno until they are finely chopped. Add to the shallots. Squeeze the lime juice over top and zest the lime rind into the mixture. Stir in the remaining ingredients (sugar, vinegar, salt and pepper) and bring to a boil. Reduce heat and simmer, stirring occasionally, until berries pop, approximately 10 minutes. Cook for an additional 5 minutes after the berries pop so that your flavours blend well.
- (5) Prepare the Jars. Put glass mason jars in the oven at 200 degree Celsius while you are cooking the chutney. Boil the jar lids in a small amount of water. Once the chutney is cooked, put the hot chutney into the hot jars. Clean off the edges of the jar to ensure that you will get a nice seal. Put the hot jar lid on top of the jar and screw the top on securely. Let the jars cool. You'll hear a pop when the jar seals. You can store the chutney until you want to use it.
- (6) Serve. This spicy wild cranberry chutney is delicious when served with crackers and cheese. You can also use it as a condiment for main dishes like turkey dinner.

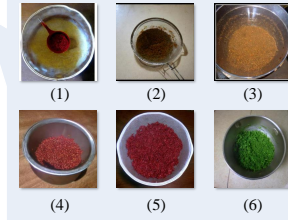
According to the above steps, can you generate images for each step?



Emu2



SEED-Llama

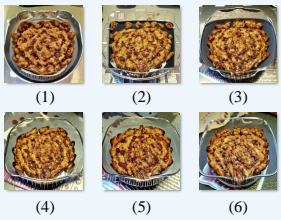


MiniGPT

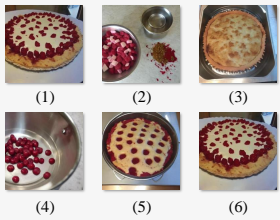


Here are step-by-step instructions without images about "Easy As Pie" Pie:

- (1) "Easy As Pie" Pie. My mom used to make a cobbler recipe with a cookie dough crust so for this pie day I decided to turn it upsidedown and make it a pie! It's so simple, if you can use a knife and turn on an oven you can make this pie!
- (2) Ingredients. Here's what you need: -roll of premade cookie dough -whatever fruit you want to put in the pie. Fresh fruit is best, but it's winter so I got a bag of frozen mixed berries which will work fine. -corn starch -optional cinnamon and sugar if you're feeling fancy. The crust is really sweet though, so you don't need much sugar. - butter or oil to grease the pie pan.
- (3) Make the Crust. Cut the roll of cookie dough into slices about one quarter inch thick. This works best if the cookie dough is frozen solid, cut it with a sharp knife and keep it in the wrapper. Arrange the cookie dough rounds in the pie pan after removing the wrapper. It's OK if there is space between them, the cookie dough will spread out as it cooks. Cook it according to the directions on the package, in this case, 10 minutes at 325 degrees. Open the oven partway through and use a fork to push the dough down if it starts getting too puffy.
- (4) Make the Filling. Mix your fruit with the corn starch and cinnamon and sugar if you're using it.
- (5) Fill Pie Shell and Bake. Put your fruit filling in the pie crust you just made and bake for 20-30 minutes at 375 degrees. Take the pie out and arrange the rest of the cookie dough on top of it and bake for 10-15 more minutes or until the cookie dough on top is golden brown.
- (6) Enjoy Your Pie! Wait a few minutes for your pie to cool and then eat it! It's great with ice cream! According to the above steps, can you generate images for each step?



Emu2



SEED-Llama



MiniGPT

Figure 7. Visualization of text-to-image sequence generation from Emu2 [36], SEED-Llama [9], and MiniGPT-5 [44], separately.

Interleaved Image-Text Content Continuation



Here are step-by-step instructions with images about Easy Homemade Khachapuri Recipe:


(1) Step 1: Cut the Dough. Add pizza stone or large baking sheet to the oven. Preheat to 450°F. On a lightly floured surface, cut the dough into 4 equal sized pieces (alternately, you can make two larger portions of cheese bread rather than 4 smaller portions). One at a time, roll the dough balls into long, thin ovals. *[Input Image 1]*

(2) Step 2: Make the Filling. Stir together the ricotta, mozzarella (or feta), whole egg, salt and pepper. *[Input Image 2]*

(3) Step 3: Fill. Add the dough ovals to a sheet of parchment paper, then add the filling to the center of each dough oval. Form into boats around the filling, twisting and pinching the edges. *[Input Image 3]*

According to the above steps, can you generate the rest steps?





(4) Step 4: Bake. Bake the khachapuri until the crusts are golden brown and the centers are puffed and set, about 10 to 15 minutes. Let cool slightly before serving.



Emu2



(4) Step 4: Bake. Place the filled dough boats on a baking sheet lined with parchment paper. Bake for 15-20 minutes, until golden brown.

(5) Step 5: Serve. Remove from the oven and serve immediately. Enjoy!



SEED-Llama



(4) Step 4: Bake. Brush the outside of the boats with the egg. Bake for 12-15 minutes, until golden brown and cooked through.



MiniGPT



Here are step-by-step instructions with images about Garlic Butter Cod Recipe:

(1) Mix the flour and spices. On a plate or shallow dish, toss the flour and spices together. *[Input Image 1]*

(2) Coat the cod fillets. Coat both sides of each fillet in the flour mixture. *[Input Image 2]*

(3) Cook the cod fillets. In a large skillet over medium heat, melt 2 tablespoons of butter. Cook the cod until it is golden brown and flaky, about 3 to 4 minutes on each side. *[Input Image 3]* Tip. Don't worry if the cod slightly falls apart. You will be deglazing the pan to make a sauce, so nothing will go to waste.

According to the above steps, can you generate the rest steps?





(4) Deglaze the pan. Deglaze the pan by adding 1 tablespoon of butter and the white wine. Scrape the bottom and sides of the pan with a wooden spoon or a whisk.



(5) Serve with a pat of garlic butter. Place a pat of butter in a bowl, and a spoon of cod on top.



Emu2




(4) Make the garlic butter sauce. Add the remaining 2 tablespoons of butter to the pan and cook until it is melted and foamy. Then, add the minced garlic and cook for 1 minute.



(5) Finish the sauce. Add the white wine and bring the sauce to a simmer. Cook for 1 minute.



(6) Serve. Serve the cod with the garlic butter sauce.



SEED-Llama



(4) Make the garlic butter sauce. Add 2 tablespoons of butter to the pan. Once it is melted, add 3 to 4 cloves of minced garlic. Cook until the garlic is translucent, about 1 to 2 minutes.



(5) Serve. Garnish with parsley and enjoy.



MiniGPT

Figure 8. Visualization of interleaved image-text content continuation from Emu2 [36], SEED-Llama [9], and MiniGPT-5 [44], separately.

Question-based Interleaved Image-Text Generation



Can you generate step-by-step instructions with images about How to Make Domino's Cheesy Bread?



- (1) Step 1. Preheat oven to 375 degrees.
- (2) Step 2. Arrange buns in baking dish and sprinkle with Parmesan cheese.
- (3) Step 3. Bake for 15-20 minutes.
- (4) Step 4. Enjoy!



Emu2



- (1) How to Make Domino's Cheesy Bread. In this instructable I will show you how to make a cheesy bread that tastes like Domino's. This is a great snack to make for your family or friends.
- (2) Ingredients: 1 cup of flour, 1/2 cup of warm water, 1/2 cup of milk..... 1/2 teaspoon of oregano.
- (3) Mix the Dough. First, in a bowl mix together the flour, salt, baking powder, sugar, garlic powder, onion powder, and oregano. Next, add the egg and mix until it is well combined. Finally, add the milk and mix until it is smooth.
- (4) Make the Cheese Sauce. In a bowl, mix together the butter, milk, salt, and cheese. Heat the mixture in the microwave for 1 minute or until the cheese is melted.
- (5) Add the Cheese Sauce to the Dough. Add the cheese sauce to the dough and mix until it is well combined.
- (6) Roll Out the Dough and Cut the Bread. Now, roll out the dough and cut it into 12 slices.
- (7) Bake the Bread. Preheat the oven to 350 degrees Fahrenheit. Place the bread slices on a baking sheet and bake for 15 minutes.
- (8) Serve and Enjoy! Serve the bread with some butter and enjoy.



SEED-Llama



- (1) Prepare the dough. It's best to do this several hours in advance of baking. Punch down the dough, then put it into a bowl. Cover it with the dough, then put it in the refrigerator.
- (2) Clean and slice the bread. Slice the dough into 1/2 inch strips. Remove the seeds from the bread and chop the seeds with a knife.
- (3) Roll the dough. Divide the dough into 1/2 inch balls. Then, using your hands, flatten the dough to make a thin disk. Place the dough in the palm of your hand and use your thumb and first two fingers to roll the dough like a ball.
- (4) Place the dough on a baking sheet. Cover the baking sheet with baking paper and bake the dough until it is brown on one side. Then, turn it over and bake it until it's browned on the other side.
- (5) Remove the dough from the oven. Fold the dough over the baking sheet to keep it hot. Then, carefully remove it from the oven with a spatula and use your hands to flatten it into a thin, round disk.
- (6) Add your ingredients. Spread the dough with marinara sauce. Then, add your favorite cheeses, such as Mozzarella, Cheddar, Parmesan, and Asiago. Then, add your favorite ingredients, such as diced tomatoes, roasted red peppers, sliced black olives, and fresh basil.
- (7) Bake the bread. Take the cheese and toppings out of the oven and let the cheese cool for a few minutes. Then, carefully add the cheese and toppings back onto the dough. Top with your favorite cheese and a pinch of dried basil. Bake the bread again until the cheese is golden brown and the crust is crispy.
- (8) Serve the cheesy bread. Slice the cheesy bread and serve with a side of pasta.



MiniGPT



Can you generate step-by-step instructions with images about Planning Your Activities of How to Celebrate Your Anniversary at Home?



- (1) Keep your budget in mind when deciding what to do. Plan to spend money on activities that will make both of you happy and are meaningful to your relationship..... You might also consider paying for some activities with gift cards or points on your credit cards that you have accumulated.
- (2) Decide what food and drink you'd like to serve. You can get takeout or make food at home..... If you've never cooked a meal for your partner, now is the time to do it.
- (3) Plan a special dessert. You could make your favorite dessert, like a cake or pie..... You could also make a homemade ice cream sundae with your favorite toppings.
- (4) Think about what decorations you'd like to have..... You could also put up streamers and balloons in your favorite color.



Emu2



- (1) Plan a weekend of activities. If you're celebrating your anniversary at home, you'll have plenty of time to do all the things you love..... This can be as simple as a night of watching movies, or as elaborate as a weekend-long celebration with your family.
- (2) Plan a special meal. If you and your partner enjoy cooking, you could plan a specific meal for..... If you're celebrating your anniversary at home, you could take a break from cooking and enjoy a special meal out.
- (3) Make a list of activities. If you and your partner have a lot of things you want to do, make a list of activities you want to do..... You could also write down a list of activities you want to do together. This can help you and your partner decide how to spend your time together.



SEED-Llama



- (1) Decide on a budget. Before you start planning your anniversary celebration, you'll want to decide on your budget..... If you have children, you may want to decide how much to spend on them and on any special activities.
- (2) Make a list of special things you want to do. One good way to celebrate..... Whatever you like to do together, you can probably incorporate it into your anniversary celebration.
- (3) Make a plan for the big day. You'll want to create a schedule for the day..... You can also plan to do other things that you like to do, like reading together or going for a bike ride.
- (4) Wrap up your anniversary celebration. When your anniversary celebration is over, you can help the over..... This way, you can spend more time enjoying your anniversary celebration and less time worrying about things.
- (5) Get more ideas for your anniversary celebration. If you want to do more than just the things you originally planned, you can..... This will help you stick to your budget.



MiniGPT

Figure 9. Visualization of question-based interleaved image-Text generation from Emu2 [36], SEED-Llama [9], and MiniGPT-5 [44], separately.

