Muffin or Chihuahua?
Challenging Large Vision-Language Models with Multipanel VQA

Anonymous ACL submission

Abstract

Multipanel images, commonly seen as web screenshots, posters, etc., pervade our daily lives. These images, characterized by their composition of multiple subfigures in distinct layouts, effectively convey information to people. Toward building advanced multimodal AI applications, such as agents that understand complex scenes and navigate through webpages, the skill of multipanel visual reasoning is essential, and a comprehensive evaluation of models in this regard is important. Therefore, we introduce Multipanel Visual Question Answering (MultipanelVQA), a novel benchmark comprising 6,600 triplets of questions, answers, and multipanel images that specifically challenge models in comprehending multipanel images. Our evaluation shows that questions in the MultipanelVQA benchmark pose significant challenges to the state-of-the-art Large Vision Language Models (LVLMs) tested, even though humans can attain approximately 99% accuracy on these questions. Distinctively, the MultipanelVQA benchmark features synthetically generated multipanel images specifically crafted to isolate and assess the impact of various factors, such as the layout, on LVLMs' multipanel image comprehension abilities. As a result, in addition to benchmarking the capabilities of LVLMs in understanding multipanel images, we analyze the potential causes for LVLMs' performance and offer insights for enhancement with the synthetic data. Code and data will be released.

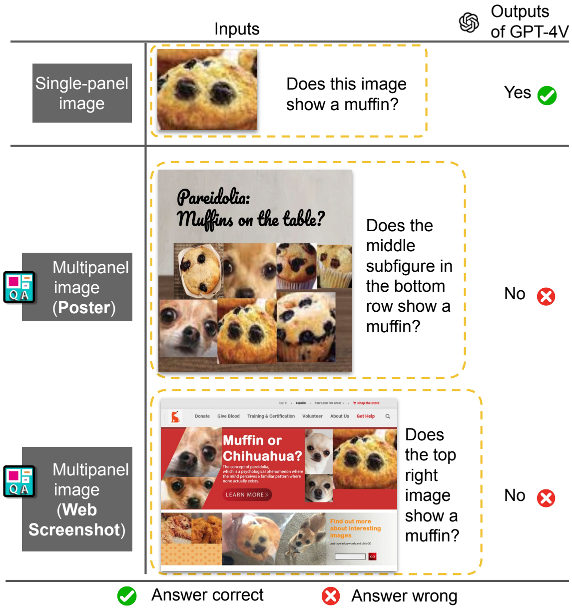


Figure 1: Examples of Single-panel *vs.* multipanel image VQA. GPT-4V distinguishes muffin and chihuahua in the single-panel image input but struggles with the same content in the multipanel image.

1 Introduction

Large Vision-Language Models (LVLMs) have become a significant leap in the integration of visual and textual data processing, enabling more nuanced understanding and generation of content that blends both visual and linguistic elements. Being trained on extensive data, advanced LVLMs (OpenAI, 2023b; Liu et al., 2023c; Ye et al., 2023b; Chen et al., 2023; Liu et al., 2023c) have shown remarkable proficiency in various tasks (e.g., image captioning and visual question answering) that require natural language understanding, visual-language grounding, visual reasoning, etc.

As LVLMs become more competent, there is a trend of establishing increasingly challenging benchmarks that are often arduous for average humans to achieve (Yue et al., 2023). However, this raises a pertinent question: Have LVLMs advanced to the stage where elementary benchmarks easily handled by average humans pose little challenge to them? To answer this question, we target multipanel images, each involving a series of subfigures. These subfigures are presented together in certain layouts, such as web screenshots capturing multiple thumbnail images and posters utilizing multipanel formats to present a cohesive narrative or argument. We observe that while humans typically find interpreting multipanel images to be a straightforward task, LVLMs struggle with this challenge when presented with the entire multipanel image as input, as shown in Figure 1.

This study aims to holistically evaluate LVLMs in understanding multipanel images. We introduce the MultipanelVQA benchmark with 6,600 triplets of multipanel images, questions and answers. The benchmark challenges models to answer each question based on the multipanel image and there are three questions with distinct types for each multipanel image: identifying common or unique contents across subfigures, pinpointing content in specific subfigures through positional descriptions, and locating subfigures via visual grounding in a multichoice format. Especially, the first type of question mainly tests the LVLMs' ability to reason about contents and the other two question types also assess the LVLMs' understanding of multipanel image layouts in addition to the content reasoning ability.

Uniquely, the multipanel images in the MultipanelVQA benchmark features a diverse mix of real-world web screenshots, posters and synthetic multipanel images, categorized into real-world data and synthetic data subsets. Unlike the real-world data that requires human annotation, the synthetic multipanel images are automatically generated by scripts with subfigures from two existed datasets. The script ensures the generated synthetic multipanel images have even distribution of various attributes such as the number of subfigures, their sizes, and the complexity of layouts, etc. As a result, based on the synthetic data, we are able to precisely isolate and assess the attributes and pinpoint their impact.

We then benchmark popular open-sourced and proprietary LVLMs on the MultipanelVQA benchmark and conduct thorough error analysis with the help of the synthetic data, which delves into the reasons behind LVLMs' difficulties in interpreting multipanel images. As a result, our main findings are 1) LVLMs are susceptible to content interference caused by the occurrence of multiple subfigures within the multipanel image. 2) The layout for subfigures has an impact on the LVLMs' performance on multipanel images. LVLMs tend to be more successful in understanding multipanel images with layouts with fewer subfigures and larger subfigure sizes. 3) LVLMs' performance can benefit from the visual text with ground truth information embedded in multipanel images.

Last but not least, we explore how adding sequential numbers to subfigure captions in multipanel images, akin to the Set-of-Mark visual prompting method (Yang et al., 2023), improves LVLMs' understanding of these images. We test LVLMs on multipanel images with and without sequential number captions for each subfigure. As a result, we observed that only GPT-4V (OpenAI, 2023b) and MiniGPT-v2 (Chen et al., 2023) show a notable improvement when the sequential number is not only embedded in the image but also explicitly mentioned in the question. In conclusion, the contributions of this study are listed as follows:

- We propose the MultipanelVQA benchmark with real-world and synthetic data that focus on evaluating the model's ability to understand the content and layout of multipanel images.
- We benchmark several open-sourced and proprietary LVLMs with the MultipanelVQA benchmark and find that all models tested face a significant challenge in interpreting multipanel images despite their success on single-panel images.
- Benefited by the synthetic data with even distributions of various multipanel image attributes in the MultipanelVQA benchmark, we conduct thorough error analysis to uncover various factors that impact the model's performance, including subfigure content, layout, background, and visual hint in multipanel images.
- Finally, we investigate the potential of adding subfigure captions in multipanel images as visual prompts to enhance the performance of LVLMs on multipanel image understanding.

2 Related Work

Large Vision Language Models The development of Large Vision Language Models (LVLMs) has been propelled by advances in large-language models (LLMs)(Chung et al., 2022; Touvron et al., 2023a,b) and vision-and-language learning(Radford et al., 2021; Li et al., 2022), merging visual comprehension with LLMs for multi-modal tasks in a zero-shot manner (Tsimpoukelli et al., 2021; Alayrac et al., 2022; Li et al., 2023b). Instruction tuning, using visual instruction data derived from open-source datasets and pre-trained LLMs, enhances LVLMs' zero-shot performance on complex tasks (Liu et al., 2023c; Zhu et al., 2023; Dai et al., 2023; Ye et al., 2023a). Further advancements include grounding and multilingual training

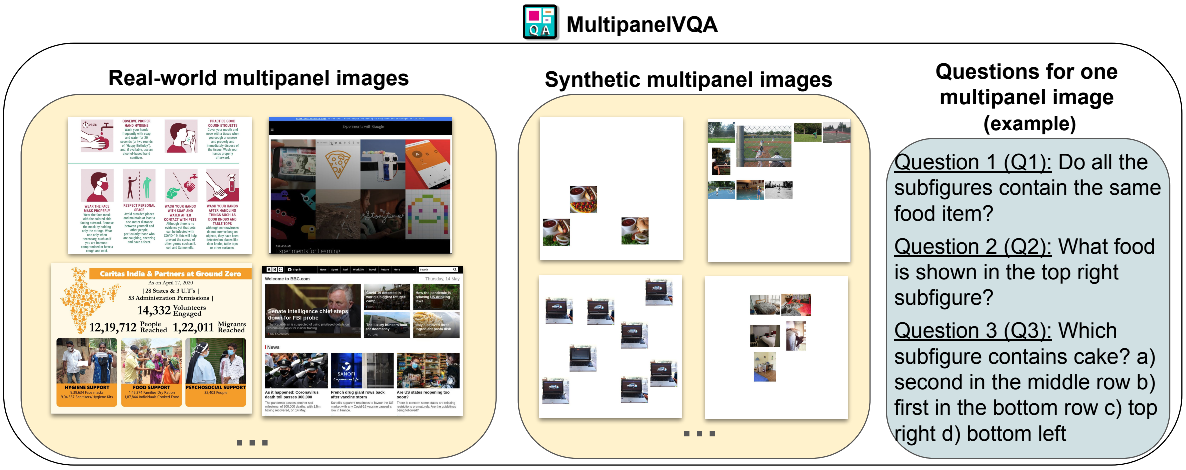


Figure 2: Overview of MultipanelVQA Data. The benchmark consists of two subsets: the synthetic data subset with artificially generated multipanel images, and the real-world data subset featuring multipanel images sourced from actual posters and web screenshots. Each image is paired with three distinct question styles, and examples of each question style are displayed on the right.

to expand LVLMs' capabilities (Chen et al., 2023; You et al., 2023; Li et al., 2023c).

Evaluations for LVLMs With the advancement of LVLMs, there's a growing need for comprehensive multi-modal benchmarks to assess their capabilities. Traditional tasks like image captioning (Chen et al., 2015; Agrawal et al., 2019) and VQA (Goyal et al., 2017; Hudson and Manning, 2019; Liu et al., 2023a), along with text recognition and knowledge-based reasoning (Marino et al., 2019; Schwenk et al., 2022; Lu et al., 2022), have been key in evaluating LVLMs. Newer benchmarks aim to assess models more holistically (Li et al., 2023a; Liu et al., 2023e; Yu et al., 2023; Cui et al., 2023). Recently, more holistic and comprehensive benchmarks have been proposed, which evaluate models' comprehensive capabilities from multiple perspectives (Li et al., 2023a; Liu et al., 2023e; Yu et al., 2023; Cui et al., 2023). Unlike former evaluation benchmarks, we propose the MultipanelVQA benchmark that not only identifies a distinguished practical challenge in real life, multipanel image understanding, but also statistically analyzes the LVLMs' capability through the synthetic data.

Synthetic Data Synthetic data, recognized for its scalability, diversity, cost-effective annotations, etc, has been widely explored for enhancing model training, especially in vision-related tasks like semantic segmentation, object detection, and image classification (Chen et al., 2019; Yuan et al., 2023; Jahanian et al., 2021; Zhou et al., 2023). Additionally, synthetic data's role extends beyond training to include model performance evaluation and analysis. Kortylewski et al. (2019) use synthetic faces to analyze neural network generalization across different poses, finding deeper architectures perform better. van Breugel et al. (2023) propose the 3S Testing framework to generate synthetic test sets that evaluate models under distributional shifts. In this work, we introduce the MultipanelVQA benchmark, enriched with synthetic data to conduct error analysis, exploring the factors influencing the performance of LVLMs on multipanel image understanding.

3 MultipanelVQA

3.1 Overview

We introduce the MultipanelVQA benchmark, consisting of multipanel images, questions, and answers, specially designed to assess the performance of LVLMs in interpreting multipanel images. As shown in Figure 2, the benchmark comprises two subsets: the real-world data subset, including actual web screenshots and posters collected by humans, and the synthetic data subset, consisting of multipanel images created by assembling individual images on blank canvases with automated scripts. As a result, the real-world subset provides realistic samples of multipanel images in everyday life, and the synthetic subset includes multipanel images with an even distribution of various attributes, including the style of the layout, number of subfigures, backgrounds, etc.

The MultipanelVQA benchmark demands that the evaluated model responds to questions linked to multipanel images, with each input consisting of a question paired with a multipanel image. As shown in Figure 2, in MultipanelVQA benchmark, there are three corresponding question-answer pairs $\{(q_{ij}, a_{ij}) | j \in [0, 2]\}$ for a multipanel image v_i. Each of the three questions features a unique style and focuses on evaluating the distinct ability of the model. Questions of the first style (Q1) assesses the model's ability to discern if any or all subfigures contain a specific object or one with unique attributes, challenging it to recognize the content of every subfigures and their spatial distinctions. The second style of question (Q2) focuses on a particular subfigure's content, while questions of the third style (Q3) features a visual grounding style in a multi-choice format requiring the model to select the positional description of the subfigure matching the given description. Notably, positional descriptions, such as "top left", exist in questions of the second and third question styles, introducing challenges due to the varying layouts of multipanel images. For example, the subfigure with a fixed position in a canvas is the topmost in one multipanel image might be the leftmost in another, depending on the arrangement of other subfigures.

3.2 Real-world Data Curation

In the real-world data subset of the MultipanelVQA, multipanel images are meticulously sourced from web screenshots in the Roboflow Website Screenshots dataset (Dwyer, 2020) and posters in task 3 of the DocVQA dataset (Mathew et al., 2021). Our data curation process begins with the manual selection of 100 images from the source, specifically chosen for their multipanel style featuring distinct subfigures. Then, for each selected image, we develop three questions. The questions are carefully designed to align with the three question styles of MultipanelVQA described in the previous section. After questions are gathered, we engage three graduate students to answer questions and validate them against the designated question types to guarantee the quality and relevance of our questions. Questions that fail validation are revised till all questions and answers are validated and collected.

3.3 Synthetic Data Curation

Generating synthetic multipanel images For the synthetic multipanel images, we use automated scripts to create multipanel images. We first generate 210 random layouts of multipanel images in different styles. Each layout holds 2 to 8 subfigures

Categories of multipanel image	Counts of image-question-answer triplets
Real-world data	300
|- Posters/Web screenshots	150/150
Synthetic data	6600
|- Original	1260
| • Subfigure quantity: 2-8	180 each
| • Subfigure source: MagicBrush/VQAv2	630/630
| • Layout Style:	
| |- Grid:	
| | same/different subfigure size	210/210
| |- Splash	210
|- Augmented:	
|- Reduced subfigure visual similarity	1260
|- Enlarged subfigure size	1260
|- With chessboard background	1260
|- With visual hint	1260

Table 1: Statistics of image-question-answer triplets in the MultipanelVQA benchmark.

and includes a predefined sequence for subfigures. As detailed in Appendix A.1, the layouts with more subfigures are populated from ones with fewer, so that when the subfigure number is increased, the positions of the existing subfigures are not changed. To generate synthetic multipanel images, we then compose single-panel images from two source datasets, MagicBrush (Zhang et al., 2023) and VQAv2 (Goyal et al., 2017), based on the layouts. Specifically, we preprocess these source datasets into sets of single-panel images with a common question and then arrange the single-panel images from the same set on a blank canvas according to the predefined layout and sequence. We provide more details about the process of multipanel image generation in Appendix A.2.

It is important to highlight that during the synthetic multipanel image curation, we filter the image sets derived from the source datasets by presenting each single-panel image within the image sets, along with the common question, to the LVLMs used in our experiments. We aim to ensure that the synthetically generated multipanel images only include subfigures that the LVLMs can accurately interpret when presented individually. This approach allows us to concentrate the evaluation squarely on the LVLMs' proficiency with multipanel images, thereby minimizing the influence of varying domain knowledge that may arise from their distinct training backgrounds.

Generating questions and answers After generating these multipanel images, we utilize GPT-4 to create questions and answers for each image, drawing on information from the source datasets.

Detailed in AppendixA.3, we design the prompt to ensure that the three questions generated for each image align with the question styles introduced in Section 3.1 consistently. For the second and third questions for each image where they target specific subfigures, human-annotated subfigure positional descriptions will be provided to GPT-4 as well. Additionally, we ensure the first subfigure added to the canvas is always the targeted subfigure, so that questions of multipanel images consisting of the same subfigure with different layouts will have similar questions that only vary on the positional description. We manually cross-validate all the questions and answers after the data curation.

Augmenting synthetic multipanel images Additionally, we uniformly augment the synthetic data subset with several variations to the multipanel images: 1) Reducing the visual similarity among subfigures in multipanel images.2) Increasing subfigure sizes while maintaining the overall multipanel image's layout. 3) Replacing the plain white background with a black and white chessboard pattern. 4) Embedding text within the images that contain ground truth information as captions for the subfigures. Please refer to Appendix A.4 for more details and examples. These augmentations enhance the complexity of the synthetic data subset of MultipanelVQA and create a test bed for comparing LVLMs' performance in interpreting multipanel images under varied conditions.

3.4 Data Statistics

Data in the MultipanelVQA benchmark comprises a substantial collection of 6,600 image-question-answer triplets, equating to unique multipanel images in two subsets: the real-world data subset, consisting of 100 multipanel images sourced from actual scenarios, and the synthetic data subset that includes a larger compilation of $2,100$ images, designed for controlled condition analysis. We detail the statistics regarding the multipanel images of MultipanelVQA in Table 1. The dataset's questions vary in length, with an average word count of 18.7. In terms of questions, 56.9% are Yes/No queries, 33.3% are multiple-choice questions, and the remainder are questions with specific categorical answers, such as identifying colors.

4 Experiments

We first evaluate six popular Large Vision-Language Models (LVLMs) on MultipanelVQA. Then, based on the evaluation result, we conduct a thorough error analysis.

	Synthetic data				Real-life data			
Models	Q1	Q2	Q3	Avg.	Q1	Q2	Q3	Avg.
Human	96.8	97.1	94.0	96.0	99.0	100.0	98.0	99.0
Random	47.2	43.5	24.4	38.4	50.0	40.0	23.0	37.7
LLaVA	75.4	60.1	28.4	54.6	70.0	56.0	49.0	58.3
MiniGPT-v2	56.4	56.3	49.9	54.2	60.0	46.0	28.0	44.7
InstructBLIP	62.6	44.1	52.4	52.7	39.0	51.0	27.0	39.0
mPLUG-Owl2	71.9	48.0	20.7	46.8	57.0	44.0	38.0	46.3
GPT-4V	**85.6**	61.0	37.8	61.5	78.0	69.0	51.0	66.0
Gemini Pro Vision	82.0	**73.9**	**59.0**	**71.6**	**81.0**	**72.0**	**64.0**	**72.3**

Table 2: Average accuracy of LVLMs on MultipanelVQA Benchmark. Q1, Q2, and Q3 represent the three question styles as introduced in Section 3.1. Two proprietary models, GPT-4V and Gemini Pro Vision, demonstrate the best overall performance. However, there is a notable gap between model and human performance.

4.1 Setup

LVLMs The LVLMs that we adopt in the evaluation include both open-source models and proprietary models with only API access. The open-source LVLMs are *(i)* LLaVA-1.5-13B (Liu et al., 2023b), *(ii)* MiniGPT4-v2 (Chen et al., 2023), *(iii)* InstructBLIP (Liu et al., 2023c) with Flan-T5 XXL (Chung et al., 2022) as the LLM backbone, and *(iv)* mPLUG-Owl2 (Ye et al., 2023b). We implement the models using their default settings and detail their supported input image resolutions in Appendix C. For proprietary models, we evaluate GPT-4V (OpenAI, 2023b) with the gpt-4-vision-preview OpenAI API during November and December of 2023 and Gemini Pro Vision(Team et al., 2023) API during January of 2024.

Evaluation In our evaluation process, we initially utilize scripts to compare the LVLM's predicted answers against the ground truth for straightforward assessments. This is particularly effective for close-ended questions like multiple-choice or yes/no questions. For cases where the LVLM's output differs from the ground truth, we employ GPT-4 (OpenAI, 2023a) as a secondary judge, assessing whether the LVLM's predicted answer, can be considered correct, especially in terms of encompassing all information present in the ground truth answer. Recent research, as cited in (Hsu et al., 2023; Hackl et al., 2023; Liu et al., 2023d), has highlighted GPT-4's capability and reliability in such evaluative roles. The details of the prompts used for this GPT-4 evaluation are provided in Ap-

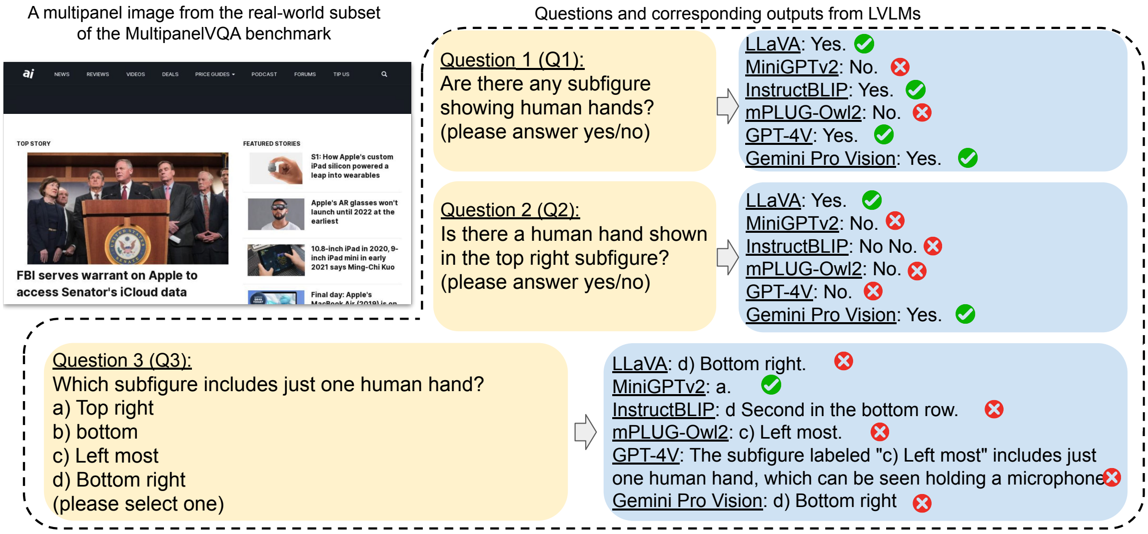


Figure 3: A sample from the real-world data subset of MultipanelVQA with outputs from models tested. The multipanel image on the left shows the characteristics of the multipanel image: complex subfigure contents and diverse subfigure layouts.

pendix D.

4.2 Main Result

We assess the performance of six leading Vision Language Models (LVLMs) using both synthetic and real-world subsets of the MultipanelVQA benchmark. Table 2 presents the accuracy of each model's output for individual questions in both subsets, alongside their averages. The result reveals that proprietary models (GPT-4V and Gemini Vision Pro) consistently outperform the other models across both subsets. However, as introduced in Section 3.2, we make sure all LVLMs tested can achieve a 100% accuracy when the subfigures are input individually, thus even the best-performing model, Gemini Vision Pro, shows an average 30% performance drop when dealing with multipanel images rather than single-panel images. Additionally, we hire human testers from both Amazon Mechanical Turk and campus to establish human performance. It's important to highlight that a significant disparity exists between the models' performances and the human-level performance, and some models even tie with the random baseline. This underscores the considerable room for improvement in current LVLMs' capabilities in handling complex multipanel image comprehension.

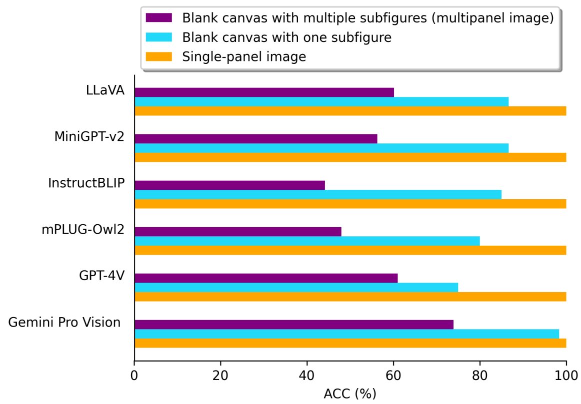


Figure 4: Model performance on questions of the second style (Q2) in the synthetic data subset when multipanel images are simplified to blank canvases, each with a targeted subfigure and then to single-panel images of the targeted subfigures, while maintaining the same input questions. The result indicates a significant vulnerability of the LVLMs to interference from adjacent subfigures.

4.3 Error Analysis

Intending to identify potential error causes, we first examine the models' outputs from the real-world data subset benchmarking results. A case study is presented in Figure 3, and we present more examples in Appendix B. Based on this example and others from the real-world data subset, we find that while the models can generate responses relevant to the posed questions, the accuracy of these answers often falls short. Based on observations, we suggest that errors in the model output primarily arise from three sources: 1) Difficulty in understanding small image sections with fewer pixels and confusion caused by neighboring subfigures in multipanel images 2) Insufficient multipanel image layout reasoning ability, and 3) Misleading factors

Interference	Content of subfigures: Visual similarity		Layout: Style		Layout: Subfigure size		Others: Background		Others: Visual hint	
Models	High	Low	Splash	Grid	Small	Large	with	without	without	with
LLaVA	53.3	58.3 (+5.0)	53.3	56.2 (+2.9)	53.3	55.3 (+2)	52.6	53.3 (+0.7)	53.3	53.7 (+0.4)
MiniGPT-v2	51.4	55.2 (+3.8)	56.4	52.6 (-3.8)	51.4	52.1 (+0.7)	54.6	51.4 (-3.2)	51.4	57.7 (+6.3)
InstructBLIP	52.4	47.9 (-5.5)	49.9	54.2 (+4.3)	52.4	51.8 (-0.6)	54.9	52.4 (-2.5)	52.4	54.9 (+2.5)
mPLUG-Owl2	47.5	43.7 (-3.8)	47.7	47.7 (+0)	47.5	45.9 (-1.6)	49.8	47.5 (-2.3)	47.5	47.3 (-0.2)
GPT-4V	59.7	62.6 (+2.9)	58.9	62.2 (+3.3)	59.7	64.3 (+4.6)	54.1	59.7 (+5.6)	59.7	66.6 (+6.9)
Gemini Pro Vision	70.2	77.9 (+7.7)	71.2	71.9 (+0.7)	70.2	71.0 (+0.8)	68.3	70.2 (+1.9)	70.2	70.7 (+0.5)

Table 3: Ablation studies of different interference factors within multipanel images, including subfigures' visual similarity, layout style, subfigure size, background, and visual hint. The columns show the accuracy of model's output in different splits of the synthetic data subset regarding various interference factors. GPT-4V and Gemini Pro Vision, being the top performers, show a marked sensitivity to these interference factors.

such as background elements and textual content within the multipanel images. However, given the complexity of real-world multipanel images, pinpointing the exact influence of each issue is difficult. Thus, we leverage the synthetic data subset of the MultipanelVQA benchmark to conduct comparative experiments isolating and evaluating the influence of distinct factors.

How susceptible are LVLMs to neighboring subfigure interference and diminished pixel detail in visual targets diminished pixel detail in specific visual targets? To evaluate the LVLMs' resilience to neighboring interference, we conduct an ablation study on the synthetic multipanel images as shown in Figure 4. Initially, for a given question targeting a subfigure within a multipanel image, we isolated the subfigure targeted by removing all others, leaving a single subfigure in the image. This modification led to improved performance across all models, suggesting their susceptibility to interference from the presence of multiple subfigures. Further, we refine the ablation to present only the target subfigure as a single-panel image input, allowing more pixels to the visual content related to the question in the image input. In this scenario, all models successfully interpreted the images, however, for most models, such improvement is less significant than the one received from the removal of neighboring subfigures. This suggests that LVLM's performance drop in understanding multipanel images is affected by both the interference from adjacent subfigures and the reduced pixel allocation to the target content but the former is more critical for most models tested.

Additionally, we explored how models' performance fluctuates with varying visual similarity of subfigures' content. From human intuition, the more similar the subfigures, the harder to distinguish the targeted subfigure. The result, depicted in Table 3, shows that except for InstrucBLIP and mPLUG-Owl2, all other models experienced a performance rise when subfigures within multipanel images are less similar.

How does LVLM's performance vary to different multipanel image layouts? We further categorize data from the synthetic data subset of MultipanelVQA based on the layout style and subfigure size, as shown in Table 3. We observe that multipanel image layout has varied influence among models. For models with better overall performance on multipanel image understanding, subfigure size and layout style play a crucial role, with larger subfigures and grid layout style generally leading to better performance. Moreover, we illustrate the impact of subfigure quantity on model performance in Figure 5, revealing a common trend where all models exhibit decreased effectiveness as the number of subfigures increases. This decline was more pronounced for MiniGPT-v2, mPLUG-Owl2 and Gemini Pro Vision.

What is the influence of background and visual hints on LVLM's multipanel image interpretation ability? Last but not least, we also investigate how other sources of interference affect the ability of LVLMs to interpret multipanel images, specifically background elements and text as visual hints. We compare the performance changes in LVLMs when presented with varying background complexities and the presence or absence of subfigure captions with ground truth information as visual hints as shown in Figure 6. As indicated in Table 3, GPT-4V and Gemini Pro Vision are the only models that show substantial improvements when the background is eliminated. However, the inclusion of visual text hint appears to enhance

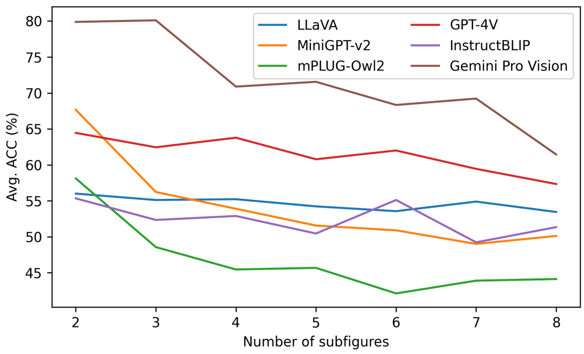


Figure 5: Impact of Subfigure Quantity on Model Performance. A common trend exists where all models exhibit declining performance as the number of subfigures increases, with varying degrees of impact.

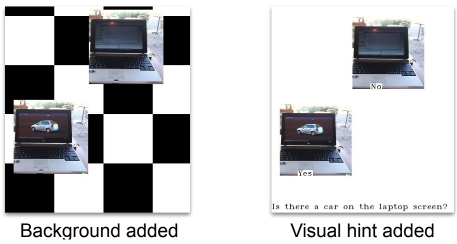


Figure 6: Demonstrations of augmented synthetic multipanel images with chessboard background (left) and embedded texts with ground truth information as visual hint (right).

the performance of nearly all models, without any detrimental effects on any model. This enhancement across all models suggests that visual text hints serve as a valuable aid in guiding the models towards better multipanel image understanding.

4.4 Influence of Adding Subfigure Captions with Sequential Numbers as Visual Prompts

Based on our findings of the visual hint's influence over the interpretative capabilities of LVLMs on multipanel images, we explore adding captions with sequential numbers for subfigures as visual prompts, akin to the Set of Mark (SoM) visual prompting method (Yang et al., 2023). We compare the model's performance on the multipanel images in the synthetic data subset with and without such subfigure captions to assess the impact. We provide a demenstration in Appendix E. Results are shown in Table 4, revealing that applying these captions with numbers as visual prompts led to little to no improvements in model performance. However, we further attempt to not only add captions with sequential numbers but also explicitly incorporate the number from the caption into the question sent to LVLMs. We find that when the number in the targeted subfigure's caption is explicitly mentioned in the input question, MiniGPT-v2, GPT-4V, and Gemini Pro Vision demonstrate a performance enhancement. This suggests that such a visual prompting method relies not only on the marks added to the input image but also on their direct integration into the query context. The result also underscores the varying nature of LVLMs' abilities to utilize visual prompts. This necessitates further exploration and development of tailored strategies for effectively integrating visual prompts into different LVLMs.

Models	**Original synthetic multipanel images**	**Add captions for subfigures**	**Refer captions in questions**
LLaVA	59.3	57.1 (-2.2)	59.0 (-0.3)
MiniGPT-v2	48.6	56.9 (+8.3)	53.8 (+5.2)
InstructBLIP	44.3	42.9 (-1.4)	33.1 (-11.2)
mPLUG-Owl2	51.0	44.3 (-6.3)	44.5 (-6.5)
GPT-4V	60.2	54.5 (-5.7)	64.3 (+4.1)
Gemini Pro Vision	75.7	71.2 (-4.5)	77.1 (+1.4)

Table 4: LVLMs' performance on questions of the second style (Q2) for synthetic multipanel images after 1) adding subfigure captions with sequential numbers to multipanel images and 2) referring to the caption in the input question. The result shows that adding such visual prompts only benefits certain models.

5 Discussion and Conclusion

In this study, we introduce the MultipanelVQA benchmark, designed to evaluate the capability of Vision Language Models (LVLMs) in interpreting multipanel images. This benchmark, comprising both real-world and synthetic data, enables a detailed analysis of LVLMs on their multipanel image understanding abilities. Our results highlight a significant performance gap between LVLMs and humans, especially since humans achieve nearly perfect scores in this benchmark. Moreover, the synthetic data of MultipanelVQA helps isolate specific performance factors and ensures that the test images were not part of the models' training datasets. This is essential for large-scale LVLMs with undisclosed training data. The creation method for these synthetic images is replicable, enabling ongoing generation of new test images and potentially aiding broader AI evaluation efforts. We believe this study provides valuable insights and methodologies for future AI research.

6 Limitation

Our study provides an in-depth evaluation of LVLMs on multipanel images using the proposed MultipanelVQA benchmark. The use of GPT-4 as an evaluator necessitated the simplification of questions to primarily yes/no or short-answer formats to allow for automated non-human evaluation. This constraint potentially limits the assessment's depth and we leave the development of evaluation with more complex questions for future research. Additionally, the synthetic data, although crucial for statistical analysis, faces challenges due to the very poor performance of some models that are close to the random baseline. The extreme underperformance of those models restricts our error analysis, as it is difficult to derive meaningful conclusions from such low accuracy levels.

References

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. nocaps: novel object captioning at scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8948–8957.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2:3.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: Large language model as a unified interface for vision-language multi-task learning. *arXiv:2310.09478*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. 2019. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1841–1850.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. 2023. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287*.

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

Brad Dwyer. 2020. Website screenshots dataset.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Veronika Hackl, Alexandra Elena Müller, Michael Granitzer, and Maximilian Sailer. 2023. Is gpt-4 a reliable rater? evaluating consistency in gpt-4 text ratings. *arXiv preprint arXiv:2308.02575*.

Ting-Yao Hsu, Chieh-Yang Huang, Ryan Rossi, Sungchul Kim, C Lee Giles, and Ting-Hao K Huang. 2023. Gpt-4 as an effective zero-shot evaluator for scientific figure captions. *arXiv preprint arXiv:2310.15405*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Ali Jahanian, Xavier Puig, Yonglong Tian, and Phillip Isola. 2021. Generative models as a data source for multiview representation learning. *arXiv preprint arXiv:2106.05258*.

Adam Kortylewski, Bernhard Egger, Andreas Schneider, Thomas Gerig, Andreas Morel-Forster, and Thomas Vetter. 2019. Analyzing and reducing the damage of dataset bias to face recognition with synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. 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 *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, et al. 2023c. M ^{3} it: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*.

Fangyu Liu, Guy Edward Toh Emerson, and Nigel Collier. 2023a. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

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

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023e. Mmbench: Is your multi-modal model an all-around player? *arXiv:2307.06281*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

OpenAI. 2023a. Gpt-4 technical report. *Technical report*.

OpenAI. 2023b. Gpt-4v(ision) technical work and authors. *Technical report*.

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 *International conference on machine learning*, pages 8748–8763. PMLR.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.

Boris van Breugel, Nabeel Seedat, Fergus Imrie, and Mihaela van der Schaar. 2023. Can you rely on your model evaluation? improving model evaluation with synthetic test data. *arXiv preprint arXiv:2310.16524*.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023a. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023b. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2023. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Jianhao Yuan, Jie Zhang, Shuyang Sun, Philip Torr, and Bo Zhao. 2023. Real-fake: Effective training data synthesis through distribution matching. *arXiv preprint arXiv:2310.10402*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*.

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

Yongchao Zhou, Hshmat Sahak, and Jimmy Ba. 2023. Training on thin air: Improve image classification with generated data. *arXiv preprint arXiv:2305.15316*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Synthetic Data Generation Details

A.1 Layout Generation

To generate synthetic multipanel images automatically, we first develop scripts to generate random layouts for subfigures in multipanel images. There are two scripts, generating layouts in splashed and grid style respectively, where splashed style has subfigures scattered in the canvas and grid style has the subfigures tightly arranged in the canvas. We provide examples in Figure 10. Both scripts generate the layout by sequentially determining the position of maximum 8 subfigures in a 1000×1000 pixels blank canvas, where there is a random selector selecting the position and size for the next subfigure from all possible candidate positions after the last subfigure is determined. Every time a new subfigure position is determined, a new layout is generated, so the number of subfigure in the layout ranges from 2 to 8. At the same time, a subfigure sequence is recorded based on the order that their position is determined in the layout.

To generate different layout styles, each script has different rules of selecting candidate positions and the size of the next subfigures. Specifically, to generate splashed style layouts, the candidate position of the next subfigure can be anywhere in the canvas as long as it is not overlapped with existing ones and the size of the subfigure is the same within the same layout, which is randomly chosen in the range of $[180, 220]$ pixels. On the other hand, for grid style layouts, the candidate positions are restricted to be either in the same row or column as the previously determined subfigure's position. Additionally, the size of the next subfigure will be either the same as the predetermined size in the range of $[180, 220]$ pixels, or twice as large as the predetermined size. As a result, the grid style layouts we randomly generated include two layouts with all subfigures in the same size and another two layouts with different size subfigures.

A.2 Multipanel Image Generation

In order to generate multipanel images, each with a consistent source, we first preprocess both source datasets, MagicBrush (Zhang et al., 2023) and VQAv2 (Goyal et al., 2017), unifying the formats of the two source datasets to be sets of images with the same question. Specifically, for MagicBrush where there are originally sets of images, each sharing a common image as an image editing source, we create a template-based question asking about

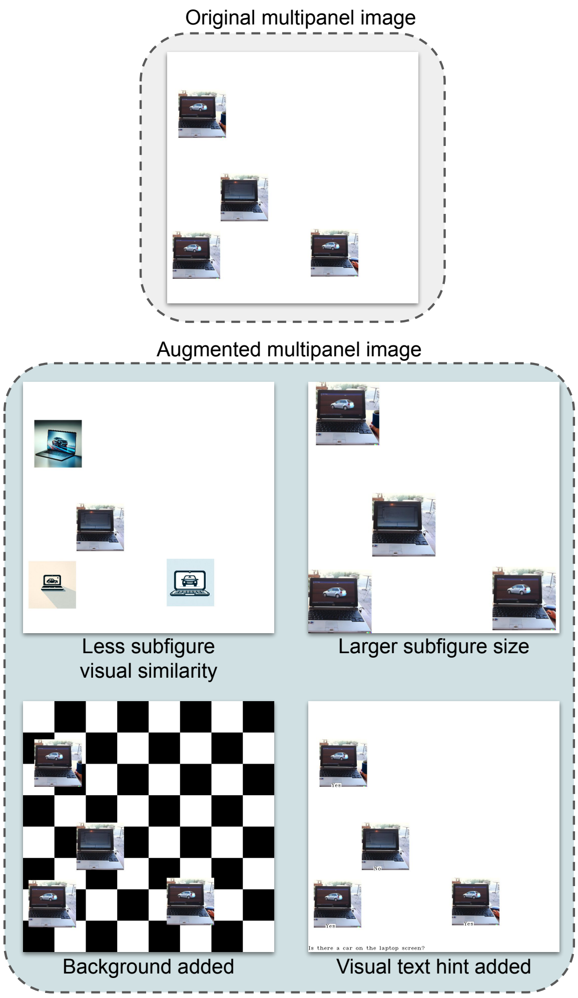


Figure 7: Examples of augmentations to synthetic multipanel images.

the visual component being edited for every image set; for VQAv2, we gather images with the same question in the dataset. We show example sets of the pre-processed source datasets in Figure 11.

Then, based on the aforementioned layouts for synthetic multipanel images and the sequence of the subfigure in the layout, we select images from the same image set in the source dataset and add them to a blank canvas. In this process, we make sure the selected images for every multipanel image include only one image with a unique answer, and we place it at the first in the sequence. Additionally, we use each image set to fill all layouts we generate, which ensures independent distributions of the subfigure content and layout.

We illustrate this process in Figure 12, where every time a new image is added to the blank canvas, a new synthetic multipanel image is created.

A.3 Question-Answer Generation

We prompt GPT-4 to generate three questions in three distinct styles and corresponding answers for each multipanel image, given the fact that all subfigures in a synthetic multipanel image come from the same image set in the source dataset and share a common question. The first question asks if all or any subfigure have a specific object or object attribute which is mentioned in the common question of the image set. The second and third will focus on the content of a specific subfigure, which is the one with a unique answer to the common question shared in the image set. The prompt, shown in Table 7 includes detailed instructions for how to generate the question-answer pairs while requiring information about the multipanel image which consists the subfigure numbers, the common question for the subfigures, the answer of the target subfigure to the common question and the positional description of the target subfigure which we manually annotate the positional description for each subfigure in advance.

A.4 Augmentation of the Synthetic Data Subset

We augment the synthetic data subset of the MultipanelVQA benchmark to enable a more comprehensive evaluation of LVLMs performance on multipanel image understanding. The augmentation is done by involving new multipanel images that are altered from the original version in four different ways while keeping the corresponding questions and answers the same. First, we reduce the visual similarity among subfigures in multipanel images by generating new subfigures to replace the original ones. Since the original subfigures in each multipanel image come from the same image set of the source dataset, they share a visual similarity as they have a common question, and many even have the same answer to the common question. In order to reduce this similarity while keeping the questions and answers for the multi-panel image unaffected, we prompt DALL·E 3 (Betker et al., 2023) to generate various images that do not incur the same answer to the common question as the target subfigure and then replace the subfigures except the target subfigure with these newly generated images. As shown in Figure 7, in this way, subfigures in multipanel images, especially those based on MagicBrush (Zhang et al., 2023) dataset, become less similar to each other visually. Second,

Models	Input image resolution
LLaVA	336
MiniGPT-v2	224
InstructBLIP	224
mPLUG-Owl	224

Table 5: Supported input image resolutions of tested LVLMs.

we increase the subfigure size within the multipanel images by first removing some edge space for the multipanel image while keeping the ratio of height and width and then resizing the image to the original size. Third, we add a background with black and white chessboard patterns to every synthetic multipanel image, introducing a more complex visual backdrop. Last, we embed texts to the multipanel image, where these texts include the common question and the corresponding answers of each subfigure.

B Samples of Model Outputs on Real-world Multipanel Images

We show some more real-world multipanel images of web screenshots and posters along with model outputs in Figure 8. Additionally, there are two examples from the synthetic data subset in Figure 9.

C Supported Input Image Resolutions of Tested LVLMs

We show the supported input image resolutions of four tested open-sourced LVLMs in Figure 5. Except for LLaVA takes input images in size of 336×336, all others have the input image limited to a size of 224×224. However, as illustrated in Figure 4, despite supporting higher-resolution input images, LLaVA's performance on simplified multipanel images—reduced to blank canvases with a single subfigure—is comparable to other evaluated open-source LVLMs. This indicates that variations in input image resolution do not significantly impact the comparative analysis of multipanel image comprehension among the tested LVLMs.

D GPT-4 as Evaluator

Given the output of LVLMs with the question and multipanel image as input, we prompt GPT-4 to judge if the output is a correct answer. The prompt is shown in Table 6, where the question, model's output and corresponding ground truth are inserted. If GPT-4's output is yes, we regard the model's output as correct and vice versa.

E Examples of Subfigure Captions with Sequential Numbers as Visual Prompts

We experiment with adding captions to subfigures in the synthetic data subset of MultipanelVQA as a visual prompting method similar to the Set of Mark (SoM) visual prompting method (Yang et al., 2023). The caption we add to the subfigures includes sequential numbers, as shown in Figure 13i. Besides changing the multipanel images with subfigure captions, we also modify the corresponding questions to refer to the subfigure caption explicitly, as shown in Figure 13ii.

Prompt:	For question: $\{question\}$ Compare the following answers: Text 1: $\{output\}$ Text 2: $\{gt\}$ Does the first one contain all key information in the second one? (yes/no) Answer:

(a)

Prompt:	For question: $\{question\}$ Ground truth: $\{gt\}$ Model predicted answer: $\{output\}$ Based on the question and the ground truth answer, is the model's predicted answer correct? If multi-choice is provided, think about which choice is selected by the model, is it correct? (please answer yes/no)

(b)

Table 6: Text prompt for GPT-4 as an evaluator to judge if the output from the model $\{output\}$ is correct given the question $\{question\}$ ground truth answer $\{gt\}$. (a) shows the prompt for GPT-4 to evaluate the model output for the first and second types of question (Q1 and Q2) in MultipanelVQA. (b) shows the prompt for GPT-4 to judge the third type of question (Q3) in MultipanelVQA

Prompt:	You are asking questions about an multi-panel image composition with multiple subfigures. You will be given a description of the overall layouts of the subfigures, a common question and answers to this question for each subfigure. First ask three questions (Q1, Q2, Q3) and then generate ground truth answers (A1, A2, A3) to each question. The second question (Q1) should be the same as the common question provided but specifically targeting at one subfigure. Make sure to include specific position of the subfigure targeted. The first question (Q1) asks if all or any subfigures have the specific object/attribute mentioned in Q2. (e.g. Do all the subfigures share certain object? Is there any subfigure has a certain object?). For both answers A1 and A2, try not to refer to specific positions of subfigures and be concise. For the third question (Q3) make it a multi-choice question with a single answer based on the common question and answer. The answer (A3) should only be the subfigure targeted. Also generate a,b,c,d four choices and randomly put the correct answer in one of them, and fill the other choices with x. For the third answer (A3), only put in the label for the correct choice (a,b,c or d). Ask questions only based on the direct information you get from the provided common question and answers. At the end of each question (Q1, Q1 or Q3), indicate what kind of answer is needed for the question. (eg. please answer yes/no, please select one). Answers generated should be consice without any explanation. Your output should be in the following format: Q1: A1: Q2: A2: Q3: A3: There are $\{num\_subfigure\}$ subfigures in the image. The common question for all subfigures are: $\{com\_question\}$. The answer from the target subfigure is: $\{answer\_target\_subfigure\}$. The answer for the other subfigures are not the same as the target subfigure. Ask questions about the target subfigure located at $\{pos\_description\}$.

Table 7: Text prompts for generating questions and answers of multipanel images in the synthetic subset of MultipanelVQA benchmark. $\{num\_subfigure\}$ is the number of subfigures in the multipanel image. $\{com\_question\}$ is the common question in the image set from source datasets. $\{answer\_target\_subfigure\}$ is the answer of the target subfigure to the common question, which is different from the answer from the other subfigures selected. $\{pos\_description\}$ is the position description for the target subfigure predefined by human.

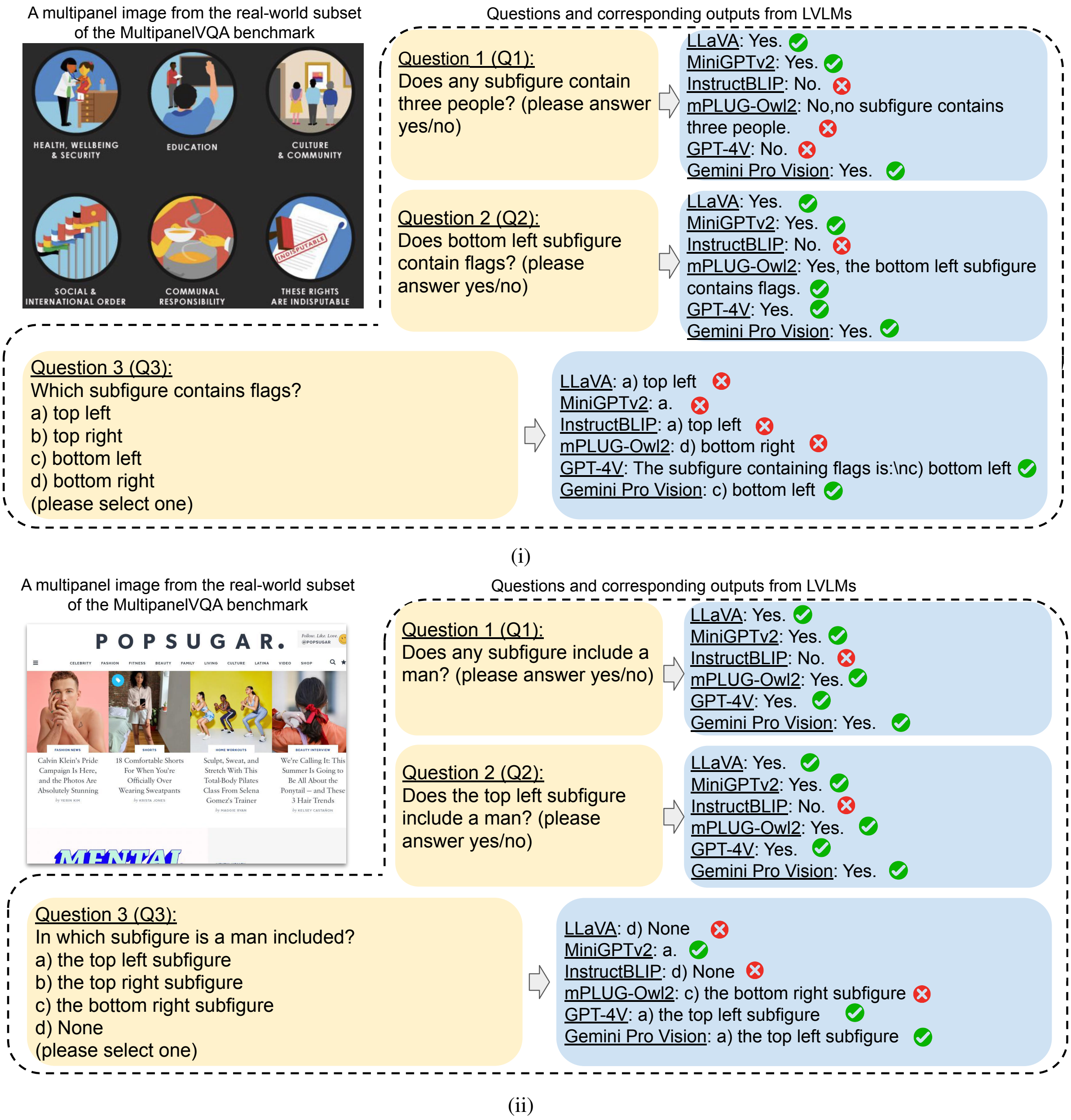
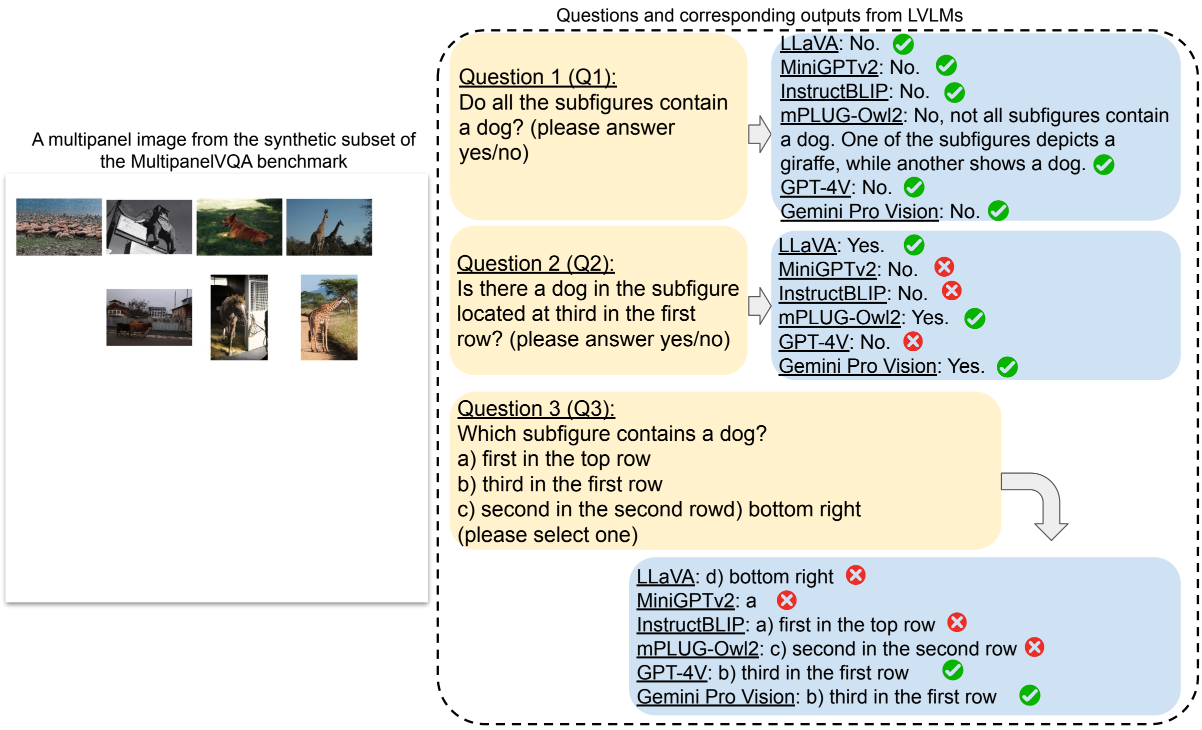
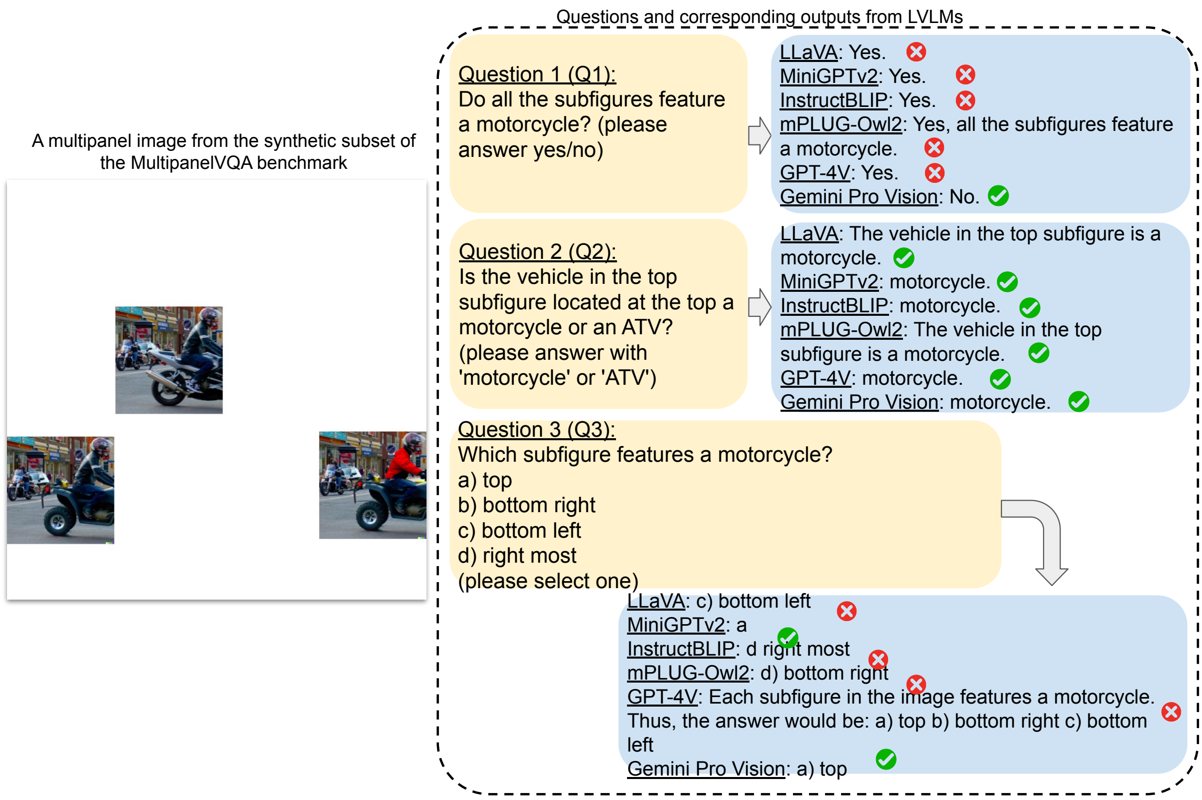


Figure 8: Samples of real-world multipanel images in the MultipanelVQA benchmark and outputs from models. (i) shows a poster multipanel image and (ii) shows a multipanel image of a web screenshot.



(i)



(ii)

Figure 9: Samples of synthetic multipanel images in the MultipanelVQA benchmark and outputs from models.

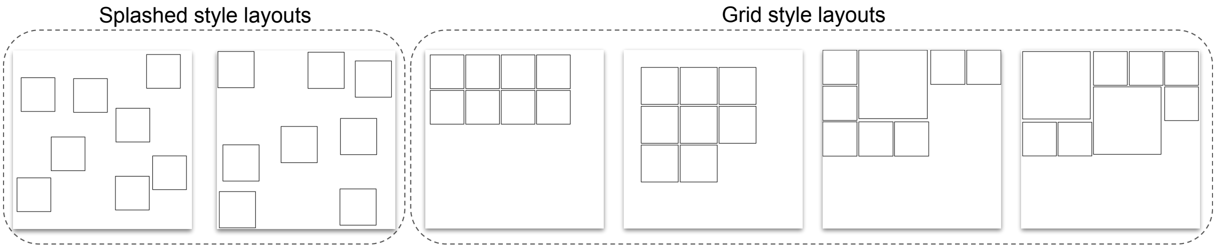


Figure 10: Examples of multipanel layouts used in the synthetic data of MultipanelVQA. The Grid style layouts include two with subfigures of the same size and another two with subfigures in two different sizes. We develop scripts to generate these layouts randomly.

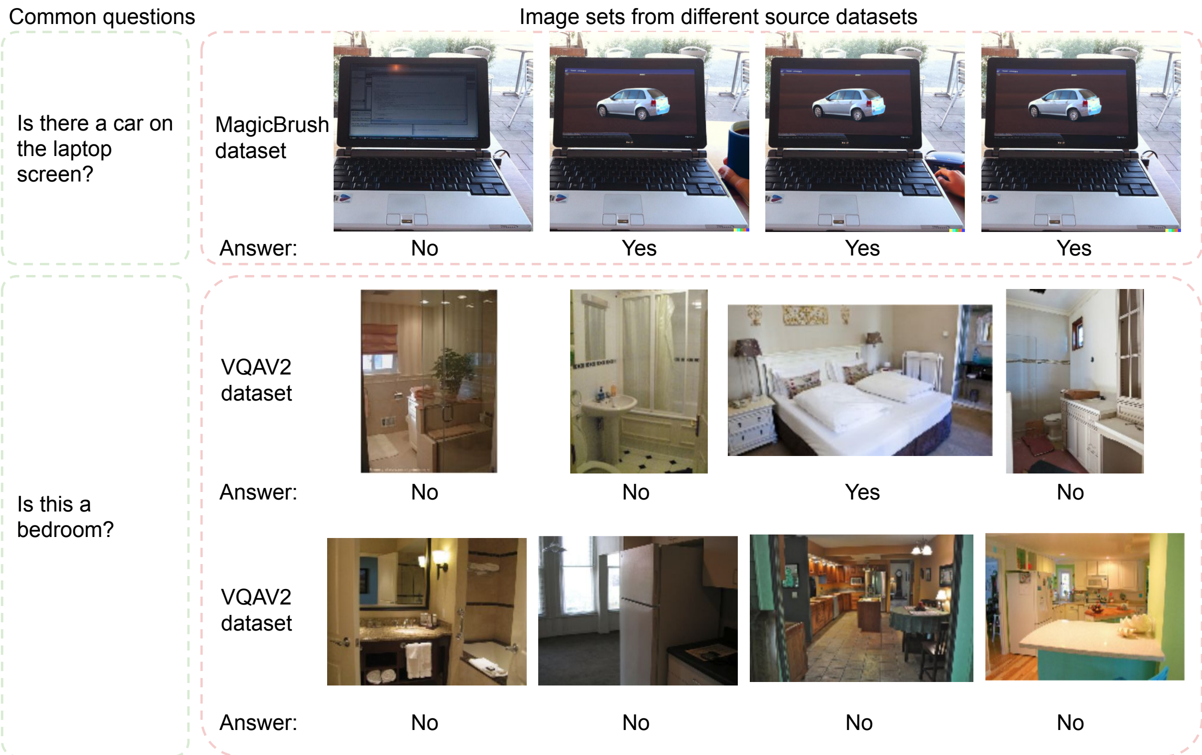


Figure 11: Examples of the image set we used from different source datasets to generate multipanel images. We prepocess two source datasets in to image sets so that images within each image set share a common question. Each image set selected includes one image that has a unique answer to the common question.

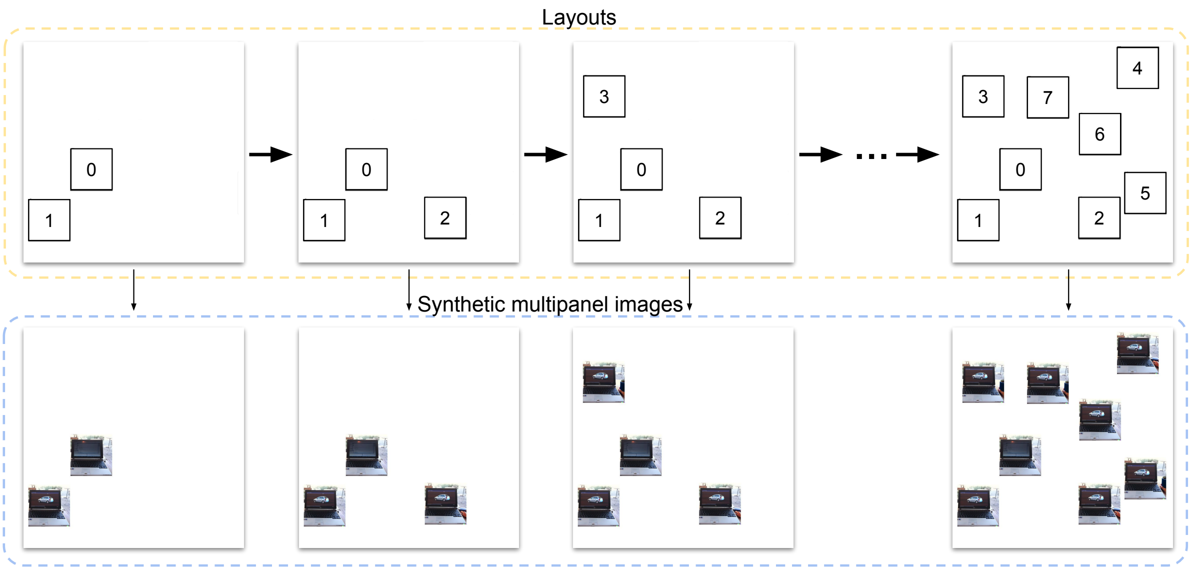


Figure 12: An example of the generation process for the layouts and synthetic multipanel images. When a new random subfigure position is determined, a new layout is formed. Based on the layouts, we position subfigures sequentially on a blank canvas according to a fixed order in each layout to create a synthetic multipanel image.

(i)

Question:
Is there a car on the laptop screen in the subfigure (Figure 0) located between the top left and bottom left? (please answer yes/no)
Answer:
No.

(ii)

Figure 13: Example for (i) a multipanel image with subfigure captions including sequential numbers and (ii) a question and answer where the question explicitly refers the subfigure caption (highlighted “Figure 0").