JourneyBench: A Challenging One-Stop Vision-Language Understanding Benchmark of Generated Images

Supplementary Material

Zhecan Wang[♦][*]   **Junzhang Liu**[♦][*]   **Chia-Wei Tang**[†]   **Hani Alomari**[†]

Anushka Sivakumar[†]   **Rui Sun**[♦]   **Wenhao Li**[♦]   **Md. Atabuzzaman**[†]   **Hammad Ayyubi**[♦]

Haoxuan You[♦]   **Alvi Ishmam**[†]   **Kai-Wei Chang**[♦]   **Shih-Fu Chang**[♦]   **Chris Thomas**[†]

[♦]Columbia University   [♦]UCLA   [†]Virginia Tech

Contents

A Project Page and Dataset Access   **3**

B Code Access   **3**

C Evaluation Procedure   **3**

   C.1 Inference Prompts 3

   C.2 MCOT/MMCOT Answer Extraction & Verification 4

   C.3 MCOT/MMCOT Solution Verification 4

   C.4 HaloQuest Answer Evaluation 5

D Detailed Experiment Results   **6**

   D.1 Experiment Results Across Five Tasks 6

   D.2 Detailed Retrieval Results 6

   D.3 Additional Multi-image VQA Results 7

   D.4 Detailed MCOT Results Across Categories 7

   D.5 Detailed Retrieval Results Across Categories 8

   D.6 Detailed Captioning Results Across Categories 9

E Annotation   **10**

   E.1 Annotation Details 10

      E.1.1 Image Filtering 10

      E.1.2 Image Captioning 10

---

[*]Equal contribution. Correspondence to: zw2627@columbia.edu

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

    E.2   Quality Assurance   11

          E.2.1   Adversarial Filtering   11

          E.2.2   Machine Focus v.s. Human focus   12

F Dataset Statistics   **14**

    F.1   General Statistics   14

    F.2   Categories Analysis   14

G Unusual Visual Scenes   **17**

    G.1   Unusual Triplet Analysis   17

    G.2   Language Prior Analysis   17

          G.2.1   LLM performance on MCOT versus ScienceQA   17

          G.2.2   LLM performance on HaloQuest versus VQA v2   17

H Image Diversity Analysis   **18**

    H.1   Image Topic Words Comparison   18

I Compututational Resources   **18**

J Comparison of JourneyBench vs. JourneyDB and WHOOPS   **18**

K Qualitative examples   **24**

L Potential Societal Impacts   **25**

M Limitations   **25**

N Personally Identifiable Information and Offensive Content   **26**

    N.1   Digital Object Identifier   26

    N.2   HaloQuest Data   26

O Future Maintenance Plan   **26**

P Terms of Usage for JourneyBench Dataset   **27**

    P.1   Ownership and Responsibility   27

    P.2   Non-commercial Research   27

    P.3   Competitive Research   27

    P.4   Restrictions on Usage   27

    P.5   Interpretation and Revision   27

    P.6   Removal of Product   28

    P.7   Licensing   28

2

A Project Page and Dataset Access

You can directly access the data via `https://journeybench.github.io/`

B Code Access

You can directly access the code via `https://github.com/JourneyBench/JourneyBench`

C Evaluation Procedure

C.1 Inference Prompts

In this section, we list the inference prompts for models to generate responses across JourneyBench tasks, including MCOT, Multi-image MCOT (MMCOT), Multi-image Case and Effect, Imaginary Image Captioning and HaloQuest (VQA with hallucination triggers).

MCOT

```
"""
You will be provided with an image and a mathematical question.
You need to solve the question with the information from the image.

Question: {$question}
"""
```

Multi-image MCOT

```
"""
You will be provided with two images and a mathematical question.
You need to solve the question with the information from the images.

Question: {$question}
"""
```

Multi-image Cause and Effect

```
"""
You will be provided with two images <image1> and <image2> and a question
querying the causal relationship between the concepts described in the
images or text.
Your final answer must be one of <image1> or <image2>.

Question: {$question}
"""
```

Imaginary Image Captioning

```
"""
Describe the unusual feature of the image in a single sentence.
"""
```

HaloQuest (VQA with Hallucination Triggers)

We use the default VQA prompt of each model. If no default VQA prompt is provided, we use the following prompt:

```
"""
Question: {$question} Answer:
"""
```

3

C.2 MCOT/MMCOT Answer Extraction & Verification

VLMs can produce not only the numerical answer but also the mathematical reasoning steps taken to arrive at the answer. Because the answer format can vary (e.g. 1/2=0.5=4/8), verifying the answer accuracy requires extra steps. Other works, for example, ScienceQA (14) use a regular expression to extract the produced answer from ChatGPT, since it consists of only multiple-choice questions. However, due to the nature of MCOT and MMCOT, distinguishing the final numerical answer from other numbers in the calculation steps can be challenging. Further, even if one prompts the VLM to produce the answer in the correct format (e.g. always express the answer in decimal on the last line), models may sometimes fail to follow the instruction or may contain a variable number of decimal points. Thus, we use Meta-Llama3-8B-Instruct (1) to first extract the answer using the prompt:

```
"""
Question: {$question}

Solution:{$reasoning_steps}.

The solution is generated by an AI model.
Identify and extract the final numeric answer from the solution.
If the answer is not explicitly stated as a number, infer it if possible.
If no numeric answer can be determined, respond with 'unknown'.
Output only the numeric answer or 'unknown'.
"""
```

Once the final numerical solution has been extracted, we then use the same model to verify the answer using the prompt:

```
"""
Question: {$question}

Predicted Answer: {$predicted_answer}

Ground Truth Answer: {$ground_truth_answer}

Does the predicted answer match the ground truth answer and directly address
the question?
If the absolute difference between their values is within 0.1, answer 'yes';
otherwise, answer 'no'. Respond only with 'yes' or 'no'.
"""
```

The verification results are reported in the form of "yes" and "no". From this, we can calculate the accuracy of the model's answers (we call this step "Answer Verification").

C.3 MCOT/MMCOT Solution Verification

We next seek to determine whether the model follows the correct logic and steps when solving the problem. We also provide step-by-step solutions in our MCOT and MMCOT annotation. To perform solution verification, we employ a Meta-Llama3-70B-Instruct (1). Essentially, we ask the language model to compare the generated solution with the ground truth provided solution and to determine whether the predicted reasoning steps follow the same approach and lead to the correct solution. We prompt Llama3-70B-Instruct using the prompt:

```
"""
Question: {$question}

Predicted Reasoning Steps: {$predicted_reasoning_steps}

Ground Truth Reasoning Steps: {$ground_truth_reasoning_steps}

Do the predicted reasoning steps follow the same approach as the ground
```

4

```
truth reasoning steps and lead to the correct solution?

Respond with 'yes' if they match, or 'no' if they differ significantly or
lead to an incorrect solution.
"""
```

C.4   HaloQuest Answer Evaluation

HaloQuest is an open-ended visual question answering dataset focusing on testing VLMs with hallu-
cination triggers. Unlike our MCOT task, HaloQuest does not ask questions requiring mathematical
problem solving skills, but instead asks general questions about the image. HaloQuest features
questions designed to trigger models to provide a hallucination response via false premise questions
(question assumes something not true in the image), visually challenging questions (answering the
question requires visual aspects of the image that are hard to see), and questions with insufficient
context to answer (asking about something not visible in the image). HaloQuest is a generalizable
dataset for future VLMs as it allows free-form answer verification, rather than requiring models to
answer multiple choice questions. We follow (21) to conduct the answer extraction and verification
process. To make the evaluation process more consistent across the five tasks in JourneyBench, we
also adopt Llama3-8B-Instruct to first extract and then verify the answer based on the raw responses,
ground-truth answers, and questions. Specifically, we used the prompt to conduct answer extraction.

```
"""
Answer extractor.

Here is my question: {$question}

Here is the response: {$response}

Can you help me extract the answer from the response to my question? Your
extracted answer should be short in one sentence.
"""
```

In addition, we used the prompt below to conduct answer verification using the LLM. That is, we had
Llama3-8B-Instruct serve as a judge by giving it the ground truth answer and the predicted answer
and asking it to determine if the predicted answer is correct given the ground truth.

```
"""
Answer verifier.

Your task is to determine if the model response is correct given the question
and ground truth response. Ensure that the model response is by the question.

If the question asks about a detail of an element that is not present in the
image, A prediction of "yes", "no" or "nothing" should be considered incorrect
because it inaccurately suggests that the element is presented in the image. The
correct prediction in such cases should acknowledge the absence of the element
in question by stating the element is not present.

If the question is about counting, then the prediction is correct only if it
matches the ground truth counts exactly.

question = {$question},
model_response = {$model_response}
groundtruth_response = {$groundtruth_response}

Please only output 'Yes' or 'No.'
"""
```

5

Model	Mean Rank	MCOT	Multi-image VQA	Captioning(C)	Text R@1	Image R@1	VQA + Hall. Trig.
GPT-4o	**2.83**	**62.18**	**56.39**	32.56			47.86
LLaVA-Next-Qwen 110B	3.16	40.43		27.18			**60.03**
LLaVA-Next-Llama3-8B	3.16	20.03		28.69			39.63
VILA-Llama3-8B	4.0	8.66	24.19	**33.79**			21.38
Mantis-Idefics2-8B	4.5	5.25	19.90	33.34			24.87
GPT-4V	4.67	49.34	48.7	11.24			44.73
BEiT3 Large-0.7B	5.16	4.10		30.90	**65.90**	56.20	34.21
Blip-2-FlanXXL-12B	6.33	3.13		26.00	63.78	**59.97**	20.56
CogVLM v2 (Llama3)-19B	8.33	8.73		30.31			38.48
InstructBLIP-Flan-T5-XXL-12B	8.5	4.31		0.46			24.83
MiniGPT4-Vicuna13B	9.0	3.73		16.21			23.39
mPLUG-Owl v2-9.2B	9.5	7.07		26.74			9.53
MiniGPT4-Llama2-7B	10.0	3.69		20.91			15.27
mPLUG-Owl-7.2B	11.83	3.19		14.68			8.05
Human		84.0	78.9	85.71			84.61
Random Basline		0	16.56		0.01	0.02	0.83

Table 1: Overview of model performance on all datasets. Captioning scores measured in CIDEr. VQA + Hall. Trig. stands for VQA + Hallucination Triggers (HaloQuest). We calculate mean rank by first ranking the model's performance on each task and taking the mean, blank cells are treated as a score of zero during ranking.

D Detailed Experiment Results

D.1 Experiment Results Across Five Tasks

In Table 1 we show a comprehensive overview of model performances on all our datasets. Note that in Table 1 we only show models that are capable of running on three or more tasks (i.e. some cross-modal retrieval models can't perform other types of tasks). We observe several surprising findings in JourneyBench. Perhaps one of the most surprising findings is that the LLaVA-Next-Qwen-110b model outperforms GPT-4o and GPT-4V significantly on the HaloQuest benchmark. This shows that GPT is significantly more prone to hallucinations than this open source model. This has implications for downstream applications where hallucination-inducing questions are likely. Users using GPT in such applications should be aware that its performance exhibits significant drops in the presence of such questions.

Note that the random baseline is higher for multi-image VQA due to the inclusion of binary cause & effect questions as part of this task. For multi-image mathematical reasoning questions, random performance is the same as MCOT.

D.2 Detailed Retrieval Results

In table 2 we include detailed cross-modal retrieval results beyond those found in our main text, including R@10 for each dataset. [2] We observe that of all models, X^2VLM-Large-590M performs quite strongly across multiple benchmarks for its size. For example, on FlickR30k, it achieves 98.8 R@1 for text retrieval and 91.8 for image retrieval, despite being more than ten times smaller than several other worse performing models (e.g. OpenCLIP-CoCa-13B, InternVL-G-14B). We observe that it also performs extremely competitively across MS-COCO and JourneyBench without distractors. However, in the presence of sample-specific distractors, it performs worse. We observe that all models are relatively close with distractors, e.g. 50s to low 60s for R@1 for image retrieval, and 60s for text retrieval. One observation is that model size may be more significant in more complex retrieval scenarios, with larger models either catching up and outperforming it on JourneyBench with distractors. This might indicate that larger models are better able to distinguish fine-grained details than X^2-VLM-Large-590M, which excels at course grained retrieval tasks.

---

[2]The ALBEF and CogVLM v2 parameter sizes in the main paper figure were labeled incorrectly and will be fixed later.

6

| | Text Retrieval | | | | | | | | | | | | | | |
| Model | Flicker30K-Full | | | MS-COCO-Full | | | MS-COCO-1K | | | JourneyBench-1K w/o distractors | | | JourneyBench-1K w/ distractors | | |
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ALBEF-210M (11)	88.5	98.5	99.2	73.96	91.8	96.0	89.1	98.3	99.6	72.3	86.1	91.78	65.36	83.75	89.13
BEiT3-674M (19)	89.5	98.8	99.4	64	86.6	92.2	81.1	96.6	98.8	74.1	87.80	92.70	65.9	86.1	90.9
BLIP2-12B (10)	92.8	99.9	99.9	80.1	94.8	97.9	91.3	99.1	99.6	**81.29**	**95.17**	**97.28**	63.78	87.76	92.46
CLIP-430M (16)	85.3	97.9	99.1	58.4	81.5	88.1	75.6	93.2	97.5	70.6	85.7	91	60.8	83.3	88.5
X^2VLM-Large-590M (23)	**98.8**	**100**	**100**	**84.4**	**96.5**	**98.5**	**93.6**	**99.5**	**99.9**	78.54	92.78	96.15	64.97	**90.47**	**94.8**
InternVL-C-13B (5)	94.7	99.6	99.9	74.9	91.3	95.2	85.34	96.86	98.84	78.22	89.21	93.61	**67.73**	86.41	91.91
InternVL-G-14B (5)	95.7	99.7	99.9	74.9	91.3	95.2	87.58	97.64	99.28	78.52	89.81	94.21	67.53	86.51	92.61
OpenCLIP-CoCa-13B (6)	92.5	99.5	99.9	66.3	86.2	91.8	75.89	93.63	97.15	70.43	85.41	89.61	60.04	83.32	87.91
	Image Retrieval														
ALBEF-210M (11)	75.9	92.6	96	54	78.99	87.18	72.28	94.18	97.54	66.12	88.65	92.15	50.02	75.46	82.56
BEiT3-674M (19)	75.94	93.34	96.66	48.9	73.2	81.8	66.4	89.5	95.2	68	90.3	94.1	56.2	79.9	85.7
BLIP2-12B (10)	89.7	98.1	98.9	63	84.2	90.2	78.78	94.92	97.74	75.77	91.66	94.12	59.97	82.48	87.17
CLIP-430M (16)	64.9	87.2	92	37.8	62.4	72.2	54.5	81.8	91	66.8	88.8	92.5	51.2	76.5	83.5
X^2VLM-Large-590M (23)	**91.8**	**98.6**	**99.5**	**67.7**	**87.5**	**92.5**	**83.32**	**96.86**	**98.6**	75.04	93.16	95.9	61.02	85	89.69
InternVL-C-13B (5)	81.7	96	98.2	54.1	77.3	84.6	71.43	91.5	96.28	75.84	93.34	96.31	62.29	83.44	89.33
InternVL-G-14B (5)	85	97	98.6	58.6	81.3	88.0	75.64	93.77	97.48	**76.8**	**93.8**	**96.4**	**63.71**	**84.84**	**90.28**
OpenCLIP-CoCa-13B (6)	80.4	95.7	97.7	51.2	74.2	82.0	59.30	85.51	92.78	65.83	86.66	91.41	48.70	72.56	80.53

Table 2: Zero-shot evaluation of retrieval tasks on different datasets along with our proposed Journey-Bench fine-grained cross-modal retrieval datasets. The best results are highlighted in bold.

| | Multi-Image VQA | | | | | | |
| Model | All | MMCOT | | | | Cause and Effect | Mantis Eval |
		All	Arithmetic Reasoning	External Knowledge	Solution Verification		
VILA-8B (12)	24.20	6.14	3.73	8.65	3.77	53.92	51.15
Idefics2-8B (9)	27.82	6.61	2.81	10.57	4.95	65.03	48.85
Mantis-Idefics2-8B (8)	19.90	3.30	3.71	2.88	7.26	49.02	57.14
Mantis-SigLIP-8B (8)	23.29	4.72	5.98	3.41	7.82	55.88	59.45
GPT-4V	48.70	32.54	32.88	**32.2**	36.31	77.06	62.67
GPT-4o	**56.39**	**41.03**	**52.04**	29.61	**43.39**	**83.33**	**73.42**
Human	78.90	71.40	86.00	55.80	-	92.00	-
Human+Internet	86.39	83.2	86.00	78.9	-	92.00	-

Table 3: Zero-shot Evaluation on Multi-Image Visual Reasoning.

D.3 Additional Multi-image VQA Results

In Table 3 we provide additional analysis of human performance on multi-image VQA. We discussed with humans performing the task why they missed certain questions. The vast majority of errors made by humans were because they lacked sufficient external knowledge about certain characters or references to the image (e.g. need to know that the Joker is a villain but Batman is a superhero) and were thus unable to figure out who or what was being referred to. To remedy this, we also granted humans access to the Internet and allowed them to search for references that they didn't recognize. We observe that after granting Internet access, the external knowledge category of MMCOT jumped significantly. This shows that our questions are highly challenging and require external knowledge to answer. We note that performance for the Cause and Effect category seems high for all models when compared to other categories, but this is because it is a binary task where random performance is 50%.

D.4 Detailed MCOT Results Across Categories

Table 4 presents detailed results of various SOTA vision-language models (VLMs) across different categories of our proposed JourneyBench MCOT dataset. Our JourneyBench MCOT dataset is

7

Model	Total	Consistency (joint accuracy)	Solution-verified	Common objects	Relevant objects with unusual properties	Irrelevant objects with unusual properties	Distractors	Occlusion	OCR	Large number	Hallucination	External knowledge
Human	84.09	44.32	82.30	88.78	84.43	84.48	82.89	81.35	96.42	81.69	82.23	76.32
LLaVA-Next-Llama3-8B	20.03	3.62	19.65	18.42	20.83	17.07	15.03	15.99	21.43	6.86	10.55	10.34
LLaVA-Next-Qwen 110B	40.43	7.46	40.28	34.58	35.00	30.89	24.85	26.32	42.26	14.29	18.81	17.24
VILA-Llama3-8B	8.66	1.71	8.31	10.69	9.38	8.13	8.28	8.91	5.36	4.57	11.47	3.45
Mantis-8B	5.25	1.07	4.67	5.54	5.21	3.25	4.60	4.05	5.26	1.14	7.80	3.45
GPT-4V	49.34	9.62	48.99	54.23	51.67	64.70	41.70	42.90	60.00	26.70	**22.89**	36.17
GPT-4o + Captioning	**62.70**	**15.35**	**62.32**	**69.16**	**70.83**	**70.73**	**54.60**	59.72	63.10	41.14	19.27	**60.34**
GPT-4o	62.18	12.15	61.97	68.90	66.45	64.70	53.90	**59.90**	**71.86**	**44.88**	13.30	58.62
InternVL-Chat-V1.5-13B	9.77	3.84	9.65	11.47	10.41	12.19	10.43	8.91	7.14	6.28	8.71	10.34
Blip-2-FlanXXL-12B	3.13	1.07	2.55	3.36	2.29	2.44	2.76	3.44	2.98	0.57	3.67	1.72
InstructBLIP-Flan-T5-XXL-12B	4.31	0.64	3.51	4.37	4.17	2.44	5.21	4.05	3.57	2.86	5.05	3.45
MiniGPT4-Vicuna-13B	3.73	0.21	3.27	3.12	4.58	0.00	5.52	3.04	4.76	1.14	12.84	3.45
MiniGPT4-Llama2-7B	3.69	0.00	3.31	3.20	3.12	1.63	5.52	2.83	4.17	1.14	9.17	1.72
mPLUG-Owl v2-9.2B	7.07	1.07	6.72	6.87	7.92	8.13	5.83	5.06	8.93	3.43	6.42	5.17
mPLUG-Owl-7.2B	3.19	3.00	2.69	3.67	2.08	2.44	5.52	2.43	1.79	0.00	5.50	0.00
CogVLM v2 (Llama3)-19B	8.73	0.21	8.23	9.44	9.17	7.31	7.97	6.07	8.92	4.00	11.00	0.00
BEiT3-674M	4.10	0.64	2.10	2.97	4.38	4.07	3.07	2.23	3.57	1.71	13.76	0.00

Table 4: Zero-shot detailed result of MCOT across categories on JourneyBench dataset. GPT-4o+Captioning indicates using GPT-4o to solve MCOT using descriptive captions of the images also generated by GPT-4o.

divided into various categories to assess the performance of different state-of-the-art vision-language models (VLMs). For the MCOT task, GPT-4o achieves the highest performance across different aspects of the dataset, obtaining an overall accuracy of 62.18%. It outperforms all other models in every category except for *Hallucination* detection, where GPT-4V demonstrates the most promising performance.

GPT-4o's superior performance extends to the relevant objects with unusual properties (66.45%) and irrelevant objects with unusual properties (64.70%) categories, indicating its adeptness at managing complex and atypical visual information. Additionally, GPT-4o shows significant strength in the OCR category (71.86%) and large numbers category (44.88%). For the external knowledge category, GPT-4o achieves the highest score (58.62%), demonstrating its proficiency in leveraging external information to enhance understanding and accuracy. Overall, GPT-4o stands out as the leading model in the MCOT task across the JourneyBench dataset, consistently outperforming other models in a wide range of categories. JourneyBench highlights GPT-4o's broad abilities to handle diverse and complex visual tasks across many different settings.

We also include the GPT-4o+Captioning result: first, we use GPT-4o to describe the image in detail, especially describing the number of each item in the image. Then, we input the question with the generated caption to GPT-4o together. However, this does not show a significant increase in the overall accuracy of the answers. The analysis in the table shows that the accuracy increased in all other categories except for the OCR and large number categories. This is possibly due to miscounting and misidentifying during the captioning phase.

D.5 Detailed Retrieval Results Across Categories

We present the performance of different state-of-the-art (SOTA) retrieval models on our proposed JourneyBench retrieval dataset in Table 5. The dataset is annotated into 11 categories, ranging from "incorrect physics rules" to "unusual attributes or accessories," to challenge the retrieval models' performance.

Overall, for the text retrieval task, InternVL-14B and OpenCLIP-CoCa generally demonstrate strong performance across most categories. In the "incorrect usage" category, BEiT3 obtains the highest R@1 score of 74.29%, which is slightly higher than InternVL-14B (70.49%) and OpenCLIP-CoCa (72.14%).

For the image retrieval task, InternVL-14B outperforms all the models across all categories of our proposed JourneyBench dataset. Across both retrieval tasks, InternVL-14B frequently appears as one of the top performers in handling diverse and complex categories within the JourneyBench dataset.

8

	Text Retrieval						
Categories	ALBEF-210M	BEiT3-674M	CLIP-430M	X2_VLM-590M	InternVL-C-13B	InternVL-G-14B	OpenCLIP-CoCa-13B
Incorrect physics rules	**69.57**	67.39	60.87	58.70	67.53	68.08	68.22
Incorrect biological rules	63.38	**67.61**	54.93	53.52	63.53	64.05	64.11
Misplacement	71.60	69.14	69.14	61.73	71.66	**71.81**	70,14
Strange animal	**73.15**	73.15	63.89	61.11	70.96	71.84	74.11
Unexpected behavior	76.47	77.65	72.55	70.98	75.20	**78.31**	70.84
Unusual food	68.75	70.83	72.92	70.83	72.02	**74.69**	71.92
Strange indoor objects	60.00	58.46	56.92	58.46	62.33	62.94	**63.77**
Strange scene	60.00	56.00	49.60	48.00	56.38	57.89	**61.00**
Unusual construction	51.52	**53.03**	43.94	45.45	52.15	52.87	52.39
Incorrect usage	60.00	**74.29**	71.43	60.00	70.97	70.49	72.14
Unusual attributes or accessories	60.70	60.95	58.46	58.21	63.28	**63.59**	55.74
	Image Retrieval						
Categories	ALBEF-210M	BEiT3-674M	CLIP-430M	X2_VLM-590M	InternVL-C-13B	InternVL-G-14B	OpenCLIP-CoCa-13B
incorrect physics rules	48.70	52.17	53.91	57.39	59.70	**60.52**	53.20
incorrect biological rules	46.20	55.21	50.70	59.72	59.78	**60.48**	53.02
misplacement	56.79	65.43	57.78	59.26	66.72	**67.44**	53.18
strange animal	55.56	60.74	47.04	66.11	63.60	**64.21**	60.47
unexpected behavior	63.37	67.37	64.16	68.63	72.28	**72.81**	61.14
unusual food	60.83	73.33	70.83	74.17	76.23	**76.97**	59.88
strange indoor objects	49.54	54.77	47.38	52.92	57.50	**58.00**	49.61
strange scene	42.64	47.52	40.08	47.52	50.76	**51.81**	43.26
unusual construction	27.58	38.18	36.97	36.67	41.55	**42.36**	28.08
incorrect usage	64.57	74.29	71.43	68.57	76.50	**77.12**	62.84
unusual attributes or accessories	48.51	52.89	49.65	53.98	57.78	**59.28**	49.13

Table 5: Zero-shot detailed results (R@1) of Retrieval across categories on our proposed Journey-Bench dataset.

Model	Overall	Incorrect physics rules	Incorrect biological rules	Misplacement	Strange animal	Unexpected behavior	Unusual food	Strange indoor objects	Strange scene	Unusual construction	Incorrect usage	Unusual attributes or accessories
Human	85.71	83.57	84.80	89.94	86.47	93.13	97.98	83.21	84.23	80.15	93.85	84.01
OpenCLIP-CoCa (Vit-L)-13B	21.59	22.32	19.25	28.24	19.43	26.87	29.65	23.16	17.89	20.71	22.40	21.93
LLaVA-Next-Llama3-8B	28.69	33.90	28.98	36.90	28.21	32.63	35.95	27.21	24.93	21.65	32.09	29.02
LLaVA-Next-Qwen110B	27.18	35.57	22.96	33.94	25.03	32.60	33.62	25.98	21.19	24.51	23.47	28.461
GPT-4o	32.56	37.05	**35.63**	**48.29**	32.98	41.97	38.52	**35.82**	25.76	20.18	**51.32**	29.06
GPT-4V	11.24	12.44	17.29	19.43	10.34	17.57	10.33	8.99	7.67	6.81	22.34	9.84
InstructBLIP-Flan-T5-XXL-12B	26.00	0.74	0.53	0.10	1.56	0.03	0.10	0.02	0.81	0.02	0.02	0.2512
MiniGPT4-Llama2-7B	20.91	27.82	23.65	24.75	23.50	25.06	21.32	22.27	16.06	15.77	29.71	18.96
MiniGPT4-Vicuna-13B	16.21	20.13	18.18	21.85	15.28	20.48	20.57	20.87	13.32	11.98	19.42	16.09
mPLUG-Owl v2-9.2B	26.74	27.39	28.03	37.34	31.62	38.46	24.33	25.46	21.24	17.77	33.80	24.68
mPLUG-Owl-7.2B	14.68	18.04	10.25	21.52	16.46	18.96	13.74	17.25	13.20	12.88	15.06	13.77
CogVLM v2 (Llama3)-19B	30.31	33.39	33.01	39.15	28.41	41.34	**41.06**	30.72	20.81	22.92	45.13	28.84
BEiT3-674M	30.90	33.25	27.08	45.27	27.12	38.39	31.59	28.49	24.64	**27.24**	28.45	31.78
Mantis_Idefics2-8B	33.34	32.88	37.47	43.91	**34.57**	**42.42**	34.41	35.46	26.67	25.18	35.62	31.09
VILA-8B	**33.79**	**39.05**	34.02	43.92	32.32	39.45	38.42	32.23	**28.12**	23.02	37.59	**34.61**

Table 6: Zero-shot detailed results (CIDEr scores) of imaginary image captioning on our proposed JourneyBench dataset. The human performance is computed by holding out one of the five annotated captions as prediction and computing the score using the rest as ground truth.

D.6 Detailed Captioning Results Across Categories

Table 6 presents the zero-shot detailed results (CIDEr scores) of various models on the imaginary image caption generation task on our proposed JourneyBench dataset. The table evaluates the models across eleven categories: incorrect physics rules, incorrect biological rules, misplacement, strange animal, unexpected behavior, unusual food, strange indoor objects, strange scene, unusual construction, incorrect usage, and unusual attributes or accessories.

To set up the benchmark performance and to illustrate the challenging nature of our proposed dataset, we also assess human performance on imaginary image caption generation. We consider this an upper bound on the captioning performance. To compute our human upper bound, we consider the set of captions for each sample. We treat each ground truth caption as a machine generated caption and use the remaining ground truth captions to compute the CIDEr score for the ground truth caption. We repeat this for every ground truth caption in each set. We find that the human CIDEr score is far higher than any machine captioning approach. This indicates to us that our captioning task is sensible (i.e. humans agree with one another on the task), but very challenging for machines given the performances shown. Human written captions achieve the highest scores in all categories of the dataset. Following human, GPT-4o, VILA and Mantis_Idefics2 models show strong performances. GPT-4o outperforms other models in misplacement (48.29%), strange indoor objects (35.82%) and incorrect usage (51.32%). VILA achieves highest scores among the models in incorrect physics rules (39.05%), strange scene (28.12%) and unusual attributes or accessories (34.61%). Mantis_idefics2

9

obtains highest scores in incorrect biological rules (37.47), strange animal (34.57%) and unexpected behavior (42.42%). However, CogVLM v2 (Llama3) outperforms all the models in unusual food category. Our results highlight the varying capabilities of different models in generating captions for unusual and complex scenarios within the JourneyBench dataset. While GPT-4o, VILA and Mantis_Idefics2 emerge as strong performers across multiple categories, the human upper bound indicates there is significant room for improvement in achieving human-like caption generation on imaginary generated images. One possible reason for this low performance is that models rely too heavily on their language biases for captioning which prevents them from describing objects or actions that are unusual.

E Annotation

E.1 Annotation Details

Q1-1: Could you understand the image content (like what is going on or depicted in the image) ?

☐ Yes

☐ No

Q1-2: Is there any obvious visual defect in the image?

☐ Yes

☐ No

If you select "**No**" to any question above, please **SKIP** all the following questions and directly jump to the next sample.

Q1-3. If you can understand the image, do you think the image is **unusual** or **fictional** (unrealistic)?

☐ Yes

☐ No

Figure 1: Annotation interface for imaginary image filtering.

E.1.1 Image Filtering

After retrieving images, human annotators filter the image set harvested using our retrieval process based on three key criteria: the images must be **unusual** or **fictional** (unrealistic), and they must also be **comprehensible**. Unusual images depict scenarios outside of everyday experiences, feature unexpected juxtapositions of objects, or include visually striking elements. Fictional images, on the other hand, present unrealistic or impossible scenes in the real world (*e.g.* an elephant standing on macaroons). However, we also enforce that the images are free of artifacts and understandable to humans to describe. This ensures a balance between creating challenging scenarios and maintaining the ability to reliably identify specific weaknesses in model reasoning or understanding. As shown in Figure 1, we assess this by directly asking annotators a set of questions, including "Can you understand the content in the image?", "Is there any obvious visual defect in the image?" and "If you can understand the image, do you think the image is unusual or fictional (unrealistic)?". We understand identifying imaginary images may be subjective, so for every image, we hire at least four Amazon Mechanical Turk (MTurk) crowd-sourced annotators to answer those questions to determine, and the 4/4 agreement is achieved in more than 72% of the cases.

E.1.2 Image Captioning

In JourneyBench, we also include a captioning task, but seek to test the models' abilities to understand and caption imaginary images. For this task, we require models to generate a single-sentence

10

Could you describe what is the most **unusual** or **fictional** (unrealistic) about this image's content?

Example

The most **unusual** or **fictional** (unrealistic) part of this image is that: an old lady wearing a purple saree is performing a difficult skateboard trick.

The most **unusual** or **fictional** (unrealistic) part of this image is that:

Figure 2: User interface for imaginary image captioning.

description of an image highlighting elements that make it imaginary. We first want to obtain the ground-truth image captions. Hence, for each collected imaginary image, we ask eight MTurk annotators to describe the most unusual or fictional part of the image in one sentence, as in Figure 2. To avoid subjective biases among annotators, those generated descriptions are further verified by another group of four experienced MTurk annotators to vote to determine whether they agree with the description. For every image, we only reserve the top five highest-voted descriptions, and each one must obtain at least two votes from the verifiers. If an image does not have five descriptions, each with at least two votes, then we believe there may not be enough agreement to determine the description, and the image is discarded.

E.2 Quality Assurance

For every step requiring annotations during our data collection process of JourneyBench, we prepare detailed instruction manuals with many examples. Given the challenging nature of our tasks, for each annotation step, we also hire at least two master annotators to supervise the annotation results for each batch to quickly verify the results by poor annotators. Defective annotations are sent back for re-correction with instructions, and annotators with quality annotation history are assigned more batches of data for annotation. Collectively, our annotators spent more than $2,200$ hours annotating JourneyBench. To help identify easy or low-quality samples, we have annotators verify the data quality of every annotated sample. To avoid human biases, we also apply adversarial models for every sample across five tasks. For instance, for MCOT questions, we leverage LLMs to guess answers and remove samples where language-only models can guess the ground-truth answers.

E.2.1 Adversarial Filtering

Filteirng via VLMs and LLMs: In order to ensure the challengeness and quality of our VLU tasks like VQA (HaloQuest), MCOT and Multi-image VQA, we inference a spectrum of VLMs of various sizes to those tasks. We filter to samples where most of VLMs can easily obtain the correct answer with high confidence scores and regard those samples as "too easy" and modify them to be more challenging or directly remove them. Additionally, to further ensure there is not shallow bias or shortcut in our data, we also apply language-only models to inference over these tasks and move the ones language-only models can score correctly.

Filtering False Positives/Negatives: Current datasets commonly used in the field often grapple with issues such as inconsistencies, false negatives, ambiguities, and more. As an illustration, Figure 3 highlights examples of false negatives within the widely-used MS COCO 5K image retrieval dataset (4), a problem largely stemming from the sampling process from the original captioning dataset. Although there have been efforts to rectify these inaccuracies (7) they have inadvertently introduced false positives, which were non-existent in the original dataset. Such examples are also depicted in Figure 3.

11

A man taking a picture of himself in a mirror.

Winter breakfast meal ready for one person at a cafe

1. A pizza sitting on top of a wooden cutting board.
2. A deep dish pizza is shown with cheese and meat toppings. ✗
3. A brick oven with logs and a uncooked pizza next to it. ✗
4. A pizza cutter is laying next to the pizza. ✗
5. A pizza cutter lying next to a well baked pizza. ✗

1. A motorcycle rider goes airborne and does tricks. ✗
2. A man that is sitting on a motorcycle in the street. ✗
3. A man is almost touching the ground while riding his motorcycle. ✗
4. A man riding on a motorcycle on the road.
5. The person on the motorcycle had a big helmet on.

Figure 3: **Top figure: false negatives in MS COCO 5K image retrieval.** These images from different data points fit the description of the same text. They are indistinguishable from the ground truth image (labeled by the green checkmark) even from the human perspective. **Bottom figure: false positives in ECCV Caption image retrieval.** A significant number of texts matched to the image by the annotation describe scenes similar to but different from the ground truth image (the red cross mark labels these captions). Evaluation results on these data points will be inaccurate.

In contrast, our retrieval dataset, despite also being sampled from our captioning dataset, primarily utilizes generated images that inherently minimize the occurrence of false negatives due to the highly randomized combination of elements within these images, a point we discussed thoroughly in the main paper. For example, in the second instance from Figure 3, the conventional dataset images involve highly related elements like "food" and "table", with a high frequency of appearing in other data points, too. In our dataset, rare combinations such as "cat" and "kimono", "CPU" and "soup", or "sander" and "donuts" (more detailed analysis in Section G.1), demonstrate a broader and more varied semantic range, with a much lower chance of having overlapping topic words among images. Finally, the prompt-based generated images on MidJourney (2) always have prompts available, which are accurate descriptors of the images, allowing us to group images by prompt to easily verify and filter false negative image-text pairs for retrieval tasks. Consequently, the likelihood of semantically similar images existing in our retrieval dataset is significantly reduced, minimizing the risk of false negatives.

E.2.2 Machine Focus v.s. Human focus

A large semantic domain for images, despite minimizing false negatives in the annotation, comes at a cost of lower retrieval difficulty, since all images/texts are highly distinct. To address this, we introduced sample-specific distractors in our retrieval dataset, as detailed in the main paper. These distractors, collected by human annotators, are both visually and semantically similar to the target images, differing only subtly to challenge the retrieval models without being misclassified as true positives.

However, the decision-making process of VL models does not always align with human judgment, as illustrated in Figure 4. The distractors collected by humans focus on certain elements like "angle" and "rocket", VL models might retrieve based on other features such as "armor" and "nature". To maintain a high level of retrieval difficulty, it is crucial to consider the perspective of VL models.

12
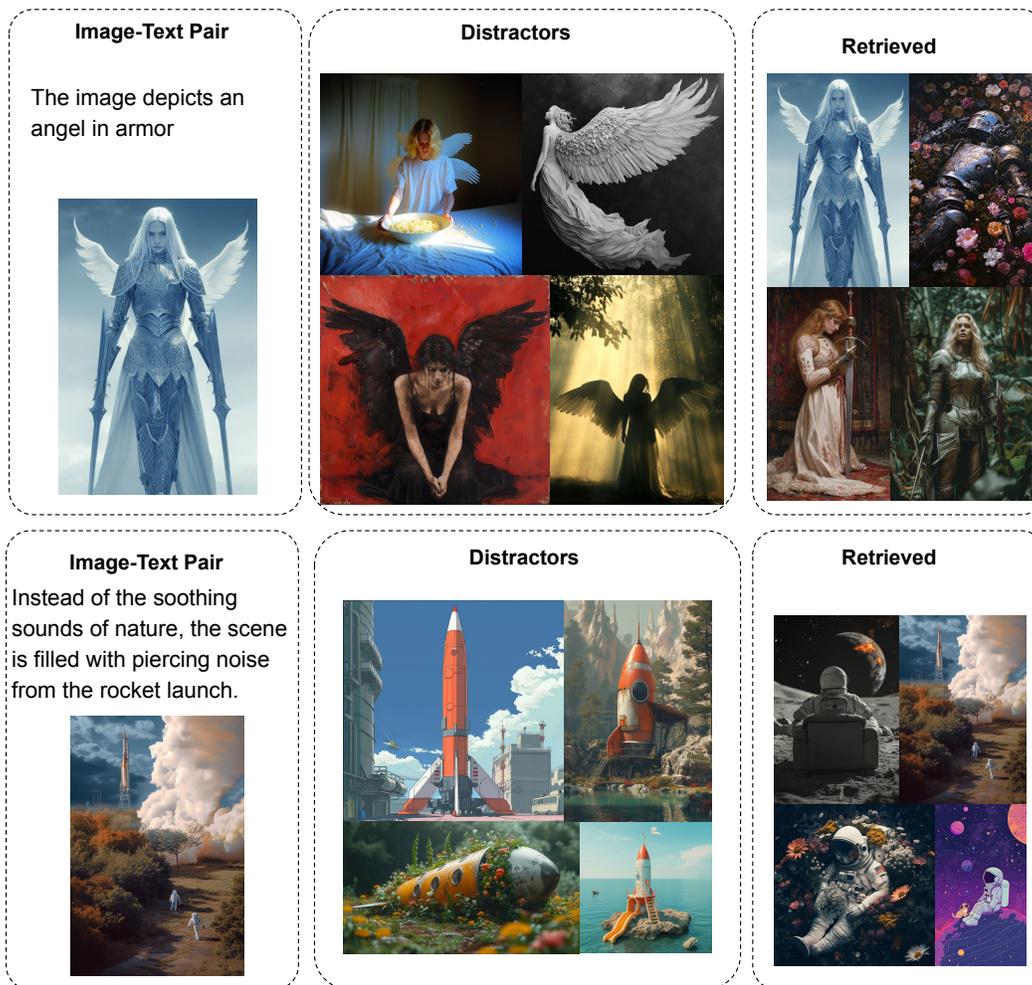
Figure 4: **Comparison between machine and human focus of images.** The distractors are collected by annotators to be semantically similar to the image. However, models sometimes do not retrieve these distractors because they focus on different aspects of the text.
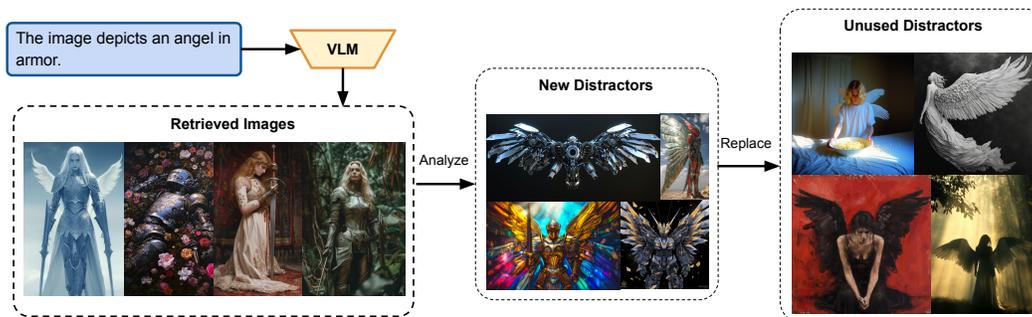
Figure 5: **One round of adversarial annotation.** The annotators analyze the retrieved images by the VL models, then collect new distractors that are closer to the models' judgment to replace the unused ones.

13

To bridge the gap between human and machine perception, we implement a multi-stage annotation process. Initially, we designate two sets of VL models — the "signal" set and the "test" set. We first evaluate the signal set models using the image retrieval dataset that includes the distractors. A distractor is deemed ineffective if none of the models retrieve it among the top five results. These ineffective distractors are then replaced based on an analysis of the top images retrieved by the models. Subsequently, we test the models on the datasets both before and after these adjustments to demonstrate the changes' effectiveness. This approach harmonizes the focal points of both humans and machines in assessing the images. Practically, we conduct two rounds of this improvement process, selecting two models each for the signal and test sets, while the remaining models are excluded from the annotation process.

F Dataset Statistics

F.1 General Statistics

Overall, JourneyBench has $13,631$ unique image-text samples across five tasks, which consist of $12,405$ unique images and $13,664$ unique text. JourneyBench includes 2,600 image-question pairs for complementary multimodal chain-of-thought, categorized into 10 fine-grained types based on visual contexts and multimodal co-referencing. All collected images in JourneyBench fall into 11 fine-grained categories based on their level of unusualness or fictionality. For multi-image VQA, there are 316 image-question pairs across three fine-grained categories. We note that this is larger than the recent multi-image VQA evaluation benchmark (217 samples) in Mantis (8). The image captioning dataset contains 1,000 images paired with 5,000 captions, with each image having five captions. For visual question answering, JourneyBench comprises 7,748 image questions, categorized into three fine-grained types of hallucination triggers. The fine-grained cross-modal retrieval task contains two subtasks. For image-to-text retrieval, there are 1,000 query images paired with 11,121 texts, averaging five positive texts (ground-truth captions) and six negative texts (sample-specific text distractors) per image. For text-to-image retrieval, there are 1,000 samples, each with five ground-truth captions, resulting in approximately 5,000 query texts against 6,323 images. Each sample has one ground-truth matching image and five negative images (sample-specific image distractors).

F.2 Categories Analysis

Imaginary Image Categories. Our imaginary image captioning dataset comprises a variety of imaginary images, classified using a set of unique categories for analysis purposes. Figure 6 displays the frequency of each category. We manually annotate each image with up to two of the 11 available categories. The diversity of scenarios challenges the models to thoroughly understand each image in order to perform effectively. Detailed examples for each category are provided in the qualitative examples section, illustrating the breadth of unusual cases that test the models' interpretive abilities.

MCOT Co-referencing Categories. As detailed in the main paper, our MCOT dataset necessitates that models reference the accompanying images to solve the math word problems presented. The questions are designed in various ways to reference images, creating diverse testing scenarios. Each data point is manually categorized to analyze the relationship between the questions and images. Figure 9 illustrates the distribution of these categories within the MCOT dataset. Additional examples from each category are available in the qualitative examples section, showcasing the range of co-referencing strategies employed in the dataset.

Multi-image Categories Our multi-image VQA dataset contains 2 tasks: multi-image MCOT and cause and effect, with multi-image MCOT further divided into two subcategories: arithmetic reasoning and external knowledge. In Figure 7 we show the percentage of each category in the dataset.

HaloQuest Categories Similar to other tasks, each HaloQuest data point is associated with a hallucination category describing the type of challenging scenario the question is testing. We show the distribution in Figure 8.

14
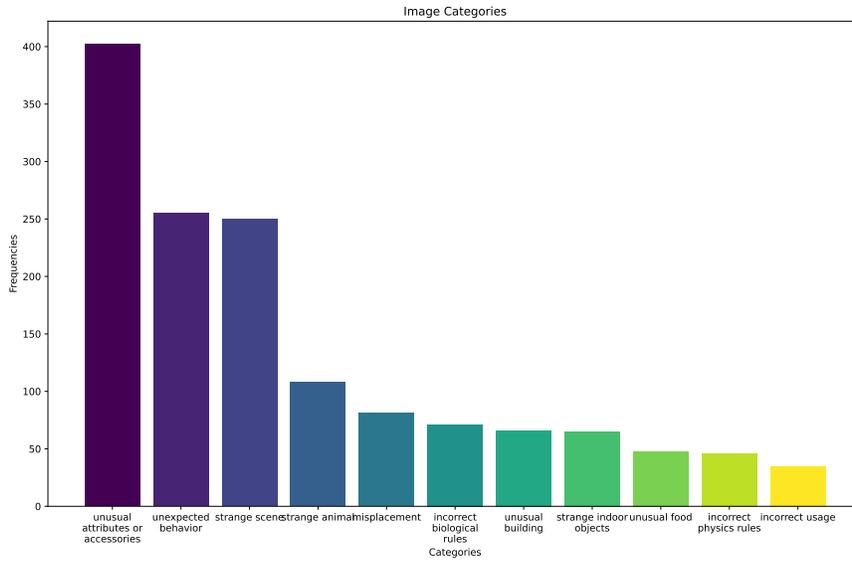
Figure 6: Frequency of categories in Imaginary Image Captioning. The categories describe the unusualness of the images.
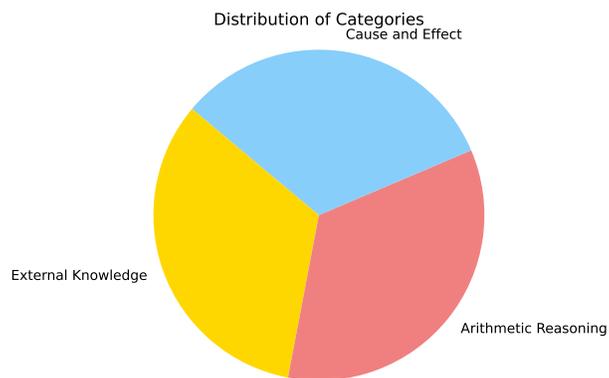
Figure 7: Frequency of categories in Multi-Image VQA.

15
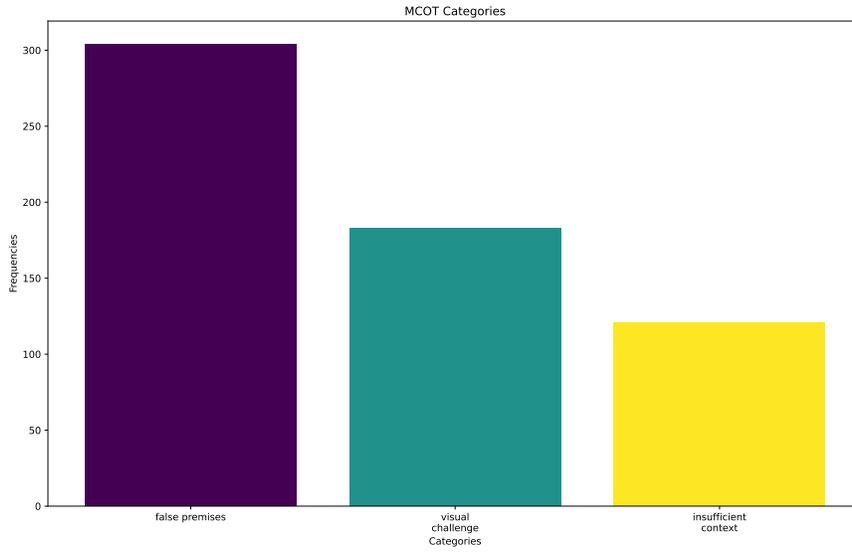
Figure 8: Frequency of categories in HaloQuest.
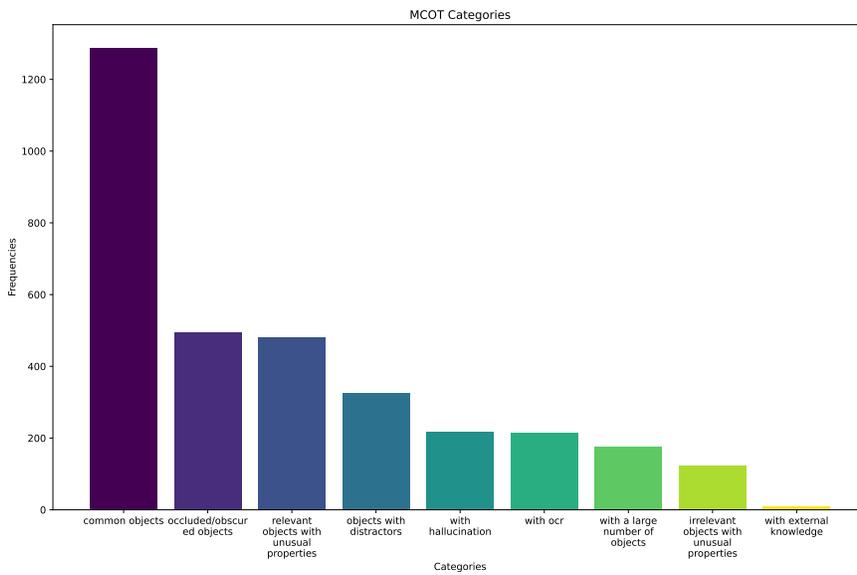
Figure 9: Frequency of categories in MCOT.

16

Dataset	Human Verify	ConceptNet
JourneyBench	8.00	6.00
COCO	72.00	68.00

Table 7: Related triplets in images. We extract triplets of subjects from images and verify their relation through human annotators and ConceptNet. JourneyBench has significantly fewer related triplets in images, indicating the unusualness of the images.

	MCOT		VQA	
Model	JourneyBench-MCOT	ScienceQA	HaloQuest	VQA v2
GPT-4o	62.18	91.04	68.10	81.84
GPT-4o (Language-only)	16.64	83.90	20.82	61.28

Table 8: Comparing the effect of removing the visual elements from datasets. MCOT and HaloQuest show a significant performance drop, indicating the strict complementing relationship between our dataset's visual and textual elements.

G Unusual Visual Scenes

G.1 Unusual Triplet Analysis

To illustrate the unusualness of JourneyBench images, we directly compared them with existing benchmarks such as MS-COCO. We randomly sampled 100 images from JourneyBench and other benchmarks, then had experienced annotators manually extract visual triplets contributing to the images' composition. These triplets, similar to unit triplets in conventional visual scene graphs, represent the visual makeup of the images. Our goal was to quantify the unusualness of these images by assessing the unusualness of the triplets based on common sense knowledge.

To evaluate the unusualness of these triplets, we used two methods. First, another group of three experienced annotators examined the triplets and voted on whether each was unusual. The label for each triplet was determined by the highest-voted option. To minimize human bias, we also employed a second approach using ConceptNet (13)[3], an external knowledge graph database. We queried ConceptNet to check if each extracted triplet existed within its database. This involved projecting the subject and object of each triplet into ConceptNet and verifying if a relationship aligned with our extracted triplet. As shown in Table 7, the majority of the triplets extracted from JourneyBench images were deemed unusual by both evaluation methods. This confirms the distinctiveness and significance of the image distribution in JourneyBench.

G.2 Language Prior Analysis

As mentioned previously, existing benchmarks consist of everyday images, which are often utilized for models' training and evaluation. This may cause existing models to develop biases of common visual compositions. Therefore, in reasoning, existing models may not fully examine the visual input information but can still resolve the task correctly based on prior knowledge. However, in edge cases in the real world, this would lead to serious application mistakes and consequences. To investigate this issue further, we directly apply language-only models to JourneyBench tasks and existing popular datasets for comparison.

G.2.1 LLM performance on MCOT versus ScienceQA

For comparison, we infer a language-only GPT-4o, over the JourneyBench MCOT dataset and another existing MCOT dataset, ScienceQA (14). From Table 8, we can observe that language-only GPT-4o can only score 16.64% on our MCOT dataset but can achieve up to 83.9% on ScienceQA.

G.2.2 LLM performance on HaloQuest versus VQA v2

We further compare language-only GPT-4o over HaloQuest versus VQA v2, a popular VQA task. From Table 8, we can observe that language-only GPT-4o can achieve much lower performance on HaloQuest compared with VQA v2.

---

[3]www.github.com/ldtoolkit/conceptnet-lite

17

Most importantly, the performance drop between GPT-4o and GPT-4o (language-only) is much larger on JourneyBecnh and much smaller on existing datasets. It is problematic that without critical visual input information, language-only models can still achieve high performances. This indicates that the underlying visual composition aligns with the models' prior knowledge or biases; thus, the visual information becomes redundant or trivial.
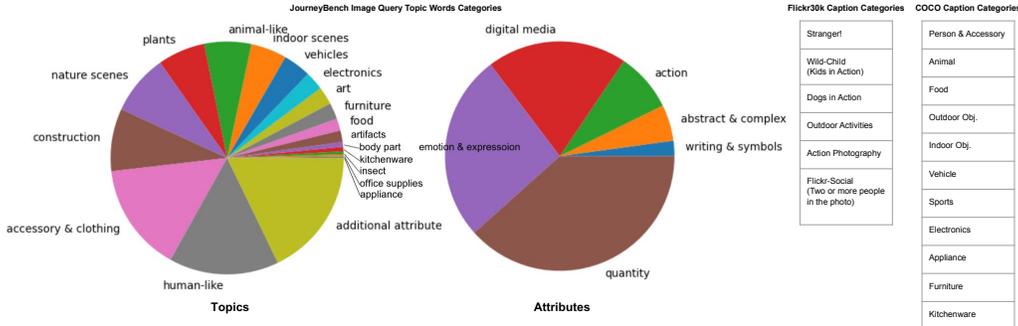
Figure 10: Topics and attributes of JourneyBench data comparing to Flickr30k and COCO Caption. Our dataset covers a much wider range of topics.

H   Image Diversity Analysis

H.1   Image Topic Words Comparison

We aim to create a VLU benchmark featuring challenging and diverse imaginary images, including unusual, abstract, and complex ones, by leveraging the advantages of prompt-based generated images. Initially, we followed the approach outlined in (3) to handcraft prompts for generating images. However, we encountered difficulties avoiding a biased image distribution and ensuring high image quality. Instead, we discovered that utilizing metadata to *retrieve* prompt-based generated images from a larger crowd-based platform provided higher quality and a more diverse distribution of images.

Thus, we developed web scraping tools to analyze metadata from Midjourney[4], which enabled us to retrieve images with a high number of views and likes. To ensure the diversity of image content, we adopted the strategy from (21), combining 17 topic words and 15 attribute words to form query words for retrieving quality images, as shown in Figure 10. This approach results in a significantly larger and more diverse set of topic words for image content compared to previous image-text datasets, which are primarily sourced from MS-COCO (4) or the Flickr platform[5].

I   Compututational Resources

To run the experiments, we utilized a cluster of A100 GPUs, A40 GPUs, and V100 GPUs. The largest and most resource intensive model we tested, LLaVA-NeXT QWEN-110B, required 4 A100 GPUs for 2 days for the MCOT task while the smallest model we tested, ALBEF-210M, required 1 V100 GPU for 1 hour for the cross-modal retrieval task. On average, depending on the task, all other models were run on 1 V100 GPU for 0-1 hour, or 1-2 A40 GPUs for 2-6 hours, or 1 A100 GPU for 1-3 hours.

J   Comparison of JourneyBench vs. JourneyDB and WHOOPS

There have been limited efforts (3; 15) to leverage generated images in VLU evaluation. These attempts have not fully exploited the controllability, convenience, and strengths of prompt-based generated images (17; 2) to address more challenging issues such as MCOT, fine-grained cross-modal retrieval (18; 24), and multi-image visual reasoning (22; 8; 20). Additionally, (3) is limited

---

[4]www.midjourney.com

[5]www.flickr.com

18

Question: Eliza's rate per hour for the first 40 hours she works each week is the value of the money bill the figure in the picture is printed on. She also receives an overtime pay of 1.2 times her regular hourly rate. If Eliza worked for 45 hours this week, how much are her earnings for this week?
Categories: External Knowledge
Answer: 4600

Question: John had a son James 60 years before he had the birthday cake in the picture. James is now twice as old as his sister Dora, who will turn 12 in 3 years. How old will John's youngest son, who was born when John was 32, in 3 years?
Categories: Common Objects, OCR
Answer: 9

Question: Dr. Hugo Grumpus and his assistant, Igor, were preparing to perform a laboratory experiment. Dr. Grumpus told Igor to gather 16 test tubes, 7 beakers, and 14 Petri dishes, and to place them all on the lab bench. By accident, Igor gathered half as many test tubes as requested. Lgor also got more Petri dishes than requested. The excess amount is the same as the petri dishes in the picture. And while he had picked up the correct number of beakers, he may lost some on the way to the lab bench. In total, the number of items Igor had placed on the lab bench was 29. How many beakers did Igor lose?
Categories: Common Objects, Hallucination
Answer: 0

Question: A tower is made out of the blue blocks and four times as many yellow blocks in the picture, and an unknown number of red blocks. If there are 32 blocks in the tower in total, how many red blocks are there?
Categories: Common Objects, Occlusion
Answer: 21

Question: Marie ordered one chicken meal that costs $12, 5 packs of milk that costs $3 each, the same number of apples as in the picture with each costing $1.50, and some boxes of pizza. Marie paid a total of $77.5. How many boxes of pizza did Marie order if each box costs $8.50?
Categories: Common Objects, Large number of objects, Occlusion
Answer: 4

Question: John has 10 hectares of a pineapple field. There are 50 times the number of pineapples in the picture per hectare. John can harvest his pineapples every 3 months. How many pineapples can John harvest within a year?
Categories: Common Objects, Distractors, Occlusion
Answer: 4000

Question: Jenna and her mother picked some apples from their apple farm. Jenna picked half as many apples as her mom. If her mom got apples 5 times the number of gas giants in the picture, how many apples did they both pick?
Categories: External Knowledge, Irrelevant Objects with Unusual Properties
Answer: 30

Question: On Monday, Sue ate 4 times as many cookies as her sister. On Tuesday, she ate twice as many cookies as her sister. On Monday her sister ate 5 times the number of cookies as the number of hearts the creature in the picture has, and 13 the next day. If 1 cookie has 200 calories, how many more calories did Sue consume than her sister?
Categories: Irrelevant objects with unusual properties, External knowledge
Answer: 11600

Figure 11: Qualitative examples of MCOT with categories.

to 500 handcrafted generated images, which not only are vulnerable to human biases in the image creation process but are much constrained in scale. On the contrary, JourneyBench has $13,631$ unique image-text samples across five tasks, which consist of $12,405$ unique images and $13,664$ unique text. Furthermore, (15) are solely annotated by a single model, GPT-3.5, and does not involve any human verification or direct annotation. Thus, it can be vulnerable to model biases and low-quality data. Differently, JourneyBench involves both human-machine-in-the-loop processes to ensure the quality and diversity of our data. Together, our annotators spent more than $2,200$ hours annotating JourneyBench.

19

Question: Toulouse has twice as many sheep as Charleston. Charleston has 4 times as many sheeps as the number of Android phones in the pictures. How many sheep do Toulouse, Charleston, and Seattle have together if Seattle has the number sheeps the same as the number of iPhones in the pictures?
Category: Arithmetic reasoning
Answer: 37

Question: In a dance class of 20 students, 20% enrolled in contemporary dance, the same number of students as people in class in the pictures enrolled in jazz dance, and the rest enrolled in hip-hop dance. What percentage of the entire students enrolled in hip-hop dance?
Category: Arithmetic reasoning
Answer: 45

Question: The Doubtfire sisters are driving home with the Siamese cats in the pictures adopted from the local animal shelter when their mother calls to inform them that their two house cats have just had kittens. She says that Patchy, the first cat, has had thrice the number of adopted cats, while Trixie, the other cat, has had 12. How many cats does the Doubtfire family now have?
Category: External knowledge
Answer: 26

Question: Kylar went to the store to buy glasses for his new apartment. One glass costs $5, but every second glass costs only 60% of the price. Kylar wants to buy the number of glasses held by the water pokemon plus twice the number of glasses held by the electric pokemon as in the pictures. How much does he need to pay for them?
Category: External knowledge
Answer: 21

Question: Uriah's book bag is getting too heavy for him. He needs to remove 15 pounds from it. His comic books weigh 1/4 pound each and his toys weigh 1/2 pound each. If he removes the Dragon Ball figures in the pictures, how many books does he need to remove?
Category: External knowledge
Answer: 54

Question: Well's mother sells watermelons, peppers, and oranges at the local store. A watermelon costs three times what each pepper costs. An orange costs 5 less than what a watermelon cost. Dillon is sent to the store to buy 5 watermelons, 20 peppers, and 5 times the number of oranges held by Tony in the picture. What's the total amount of money he will spend if each pepper costs 15$?
Category: External knowledge
Answer: 925

Question: Which one of the two images is the cause? <image1> or <image2>
Category: Cause and Effect
Answer: <image1>

Question: The man is preparing for a date. Which one of the two images shows the effect? <image1> or <image2>
Category: Cause and Effect
Answer: <image1>

Question: Which one of the two images is the effect? <image1> or <image2>
Category: Cause and Effect
Answer: <image2>

Figure 12: Qualitative examples of Multi-image VQA with categories.

20

1. A panda amusingly takes a ride on a boat.
2. A panda skillfully rowing a boat.
3. An unusual image of a panda enjoying a boat ride.
4. The image is unusual because it shows a panda on a boat.
5. A panda riding a boat.

Categories: Unexpected behavior

1. The image showcases an unconventional dish where a burrito is filled only with blueberries.
2. A burrito filled with nothing but blueberries stands out strikingly in the image.
3. The photograph presents a burrito wrapped up with blueberries instead of usual fillings.
4. The image shows a burrito rolled full of blueberries.
5. Blueberries are rolled in a burrito.

Categories: Unusual food

1. The image features a small cat hatching from a large egg.
2. In this picture, a cat is emerging from an ostrich egg.
3. The image shows a sphynx cat hatching out of an egg.
4. The picture depicts a tiny cat is coming out of an egg.
5. The image captures a cat hatching from a large egg.

Categories: Incorrect biological rules

1. In the image, a hammer is striking a hot slice of meat.
2. The picture shows a hammer is being used on a heated piece of bacon.
3. The image shows use of a hammer on a hot slice of meat.
4. The picture is strange because it shows a hammer hitting a hot piece of meat.
5. A hot slice of meat is being hitted by a hammer.

Categories: Incorrect usage

1. The image shows a wombat cradling a boombox.
2. The scene is unusual because it shows, a wombat is gripping a radio.
3. In this image, a wombat is seen holding a radio.
4. The image captures an a wombat with a radio.
5. A wombat in holding a radio up high.

Categories: Unexpected behavior

1. The image shows an unusual scene of an elephant walking underwater in the sea.
2. In the image, an elephant walks peacefully under the sea.
3. The image captures a rare and unusual sight of an elephant swimming deep in the ocean.
4. The image represents an elephant submerged under the sea.
5. An elephant is seen moving beneath the sea surface in this image.

Categories: Unexpected behavior

1. The image features a stylish black car equipped with donut wheels.
2. The black car fitted with donut-shaped wheels.
3. A black car with pink donut wheels is unusual.
4. In this picture, you can see a black car with its unusual donut wheels.
5. The image shows a black car that has donut wheels.

Categories: Unusual attributes or accessories, Incorrect usage

1. The image shows cats adorably dressed in colorful kimonos and standing on their hind legs.
2. The cats in the picture are standing upright and fashionably dressed in kimonos.
3. The cats are standing tall on their two legs and are dressed in beautiful kimonos.
4. The cats in this picture stand on their legs and are clothed in Japanese-style kimonos.
5. Cats standing and wearing traditional Japanese kimonos.

Categories: Unusual attributes or accessories, Unexpected behavior

1. The image shows tiny individuals busily repairing a motherboard.
2. People are miniaturized in the photograph, where they are working on a computer.
3. A group of worker standing on a computer.
4. A team of small workers is repairing a computer.
5. Small people working the circuits.

Categories: Unusual attributes or accessories, Strange scene

Figure 13: Qualitative examples of Imaginary Image Captioning.

21

| Text | Ground Truth Images | Distractor Images |

A cartoonish house that has sprouted legs and is strolling down a brick road.

A cat, unusually dressed in glasses, is engrossed in a scientific study.

An octopus is playing a tune on the piano.

Cat peacefully lounging on a Shiba dog.

A well-dressed dog appears deep in thought next to a piano in the picture.

The image displays an unusual scene of dogs dressed in ancient military armor.

The image displays an unusual scene of a camel appearing to hover in a vast desert.

There are four moons lighting up the night sky.

Figure 14: Qualitative examples of text-to-image retrieval with distractors.

22

Image	Ground Truth Text	Text Distractors
	1. A panda amusingly takes a ride on a boat. 2. A panda skillfully rowing a boat. 3. An unusual image of a panda enjoying a boat ride. 4. The image is unusual because it shows a panda on a boat. 5. A panda riding a boat.	1. the panda riding the boat was actually a talented animatronic, part of a new theme park attraction. 2. the panda was dressed in a sailor suit, steering his ship towards an island filled with bamboo. 3. the bamboo boat transformed beneath the panda into a shiny metallic speedboat. 4. the panda had left the bamboo forest and decided to start a career as a professional rower. 5. the panda was part of a fierce rowing competition with other animals.
	1. The image showcases an unconventional dish where a burrito is filled only with blueberries. 2. A burrito filled with nothing but blueberries stands out strikingly in the image. 3. The photograph presents a burrito wrapped up with blueberries instead of usual fillings. 4. The image shows a burrito rolled full of blueberries. 5. Blueberries are rolled in a burrito.	1. despite being a burrito, it is served with a topping of whipped cream and a cherry. 2. a side dish of vanilla ice cream suits the sweet, fruity burrito on the plate. 3. the burrito is being eaten at a fancy restaurant, known for its unique take on traditional mexican cuisine. 4. the burrito in the image is also filled with chunks of milk chocolate, making it a perfect sweet treat. 5. the blueberry burrito is surrounded by sliced strawberries for added sweetness.
	1. The image features a small cat hatching from a large egg. 2. In this picture, a cat is emerging from an ostrich egg. 3. The image shows a sphynx cat hatching out of an egg. 4. The picture depicts a tiny cat is coming out of an egg. 5. The image captures a cat hatching from a large egg.	1. The hatching cat is wearing a birthday hat. 2. the egg in the image is a robin's egg, known for its distinct blue color. 3. the cat, after hatching from the egg, starts to fly using its wings. 4. the cat has an unusual spotted pattern, similar to the texture of the egg. 5. the hatching cat has a feathery tail, resembling that of bird.
	1. In the image, a hammer is striking a hot slice of meat. 2. The picture shows a hammer is being used on a heated piece of bacon. 3. The image shows use of a hammer on a hot slice of meat. 4. The picture is strange because it shows a hammer hitting a hot piece of meat. 5. A hot slice of meat is being hitted by a hammer.	1. the image also shows a piece of bread being toasted next to the meat slice. 2. the meat slice is freezing despite its appearance. 3. it is normal to see a frying pan instead of a hammer hitting hot meat. 4. the hammer is striking an ice block instead of a meat slice. 5. the hammer is shaping a glowing piece of metal.
	1. The image shows a wombat cradling a boombox. 2. The scene is unusual because it shows, a wombat is gripping a radio. 3. In this image, a wombat is seen holding a radio. 4. The image captures an a wombat with a radio. 5. A wombat in holding a radio up high.	1. the image shows a hive mind of ants collectively holding up a radio. 2. in the picture, the wombat is pedaling a unicycle while juggling three boomboxes. 3. the picture shows the capybara playing a guitar in a band setup, which is far from holding a radio. 4. the scene shows the wombat magically levitating the boombox with its mind. 5. a kangaroo, not a wombat nor a capybara, is balancing a boombox on its tail while hopping in the australian outback.
	1. The image shows an unusual scene of an elephant walking underwater in the sea. 2. In the image, an elephant walks peacefully under the sea. 3. The image captures a rare and unusual sight of an elephant swimming deep in the ocean. 4. The image represents an elephant submerged under the sea. 5. An elephant is seen moving beneath the sea surface in this image.	1. the elephant uses special seaweed as a snorkeling mask allowing it to spend a prolonged amount of time under the sea. 2. the sea water has a sparkling azure color due to the presence of vast amounts of sapphire stones on the seabed. 3. the elephant is practicing for an underwater ballet routine, showcasing their hidden immense grace. 4. besides the elephant, there is also a group of dolphins helping him navigate underwater. 5. the elephant is searching for submerged pearls as part of a complex sea treasure hunt.
	1. The image features a stylish black car equipped with donut wheels. 2. The black car fitted with donut-shaped wheels. 3. A black car with pink donut wheels is unusual. 4. In this picture, you can see a black car with its unusual donut wheels. 5. The image shows a black car that has donut wheels.	1. the black car's donut wheels are spinning so fast, someone just got a powdered sugar dusting. 2. jelly oozes out of the donut wheels on the black car as it moves. 3. the black car is powered entirely by coffee to complement its donut wheels. 4. the black car with donut wheels is levitating above the ground. 5. the donut wheels on the black car have started to melt under the hot sun.
	1. The image shows cats adorably dressed in colorful kimonos and standing on their hind legs. 2. The cats in the picture are standing upright and fashionably dressed in kimonos. 3. The cats are standing tall on their two legs and are dressed in beautiful kimonos. 4. The cats in this picture stand on their legs and are clothed in Japanese-style kimonos. 5. Cats standing and wearing traditional Japanese kimonos.	1. the cats are discussing their secret mission.. 2. the cats are sitting on a floating disk, levitating a few inches above an ancient mosaic-studded floor. 3. the kimonos worn by the cats are woven from rare ethereal silk that can shift colors according to the wearer's moods. 4. the setting is not an ordinary room, but inside a magical japanese palace that changes its wallpaper every few minutes. 5. the cats are anthropomorphic beings, holding a high-level diplomatic meeting as representatives of their respective realms.
	1. The image shows tiny individuals busily repairing a motherboard. 2. People are miniaturized in the photograph, where they are working on a computer. 3. A group of worker standing on a computer. 4. A team of small workers is repairing a computer. 5. Small people working the circuits.	1. they work in rhythm to a peculiar tech-themed symphony. 2. The motherboard they're working on is actually part of a sentient supercomputer and sometimes communicates with the workers. 3. The workers on the computer are not just engineers but also skilled magicians who use their magical abilities to fix technical issues. 4. A curious alien is observing the group of tiny workers repairing the computer to understand human technology better.. 5. The group working on the computer has a musical conductor leading them

Figure 15: Qualitative examples of image-to-text retrieval with distractors.

23

HaloQuest

Figure 16: Qualitative examples of HaloQuest.

K Qualitative examples

Please refer to Figures 11, 12, 13, 14, 15, and 16

24

L Potential Societal Impacts

Potential Positive Impacts. The development and deployment of advanced vision-language benchmarks like JourneyBench have several potential positive societal impacts. Firstly, because JourneyBench is a one-stop vision-language benchmark with fine-grained annotations, it makes comparing the performance of different state-of-the-art AI systems easier. For example, JourneyBench exposes that GPT4o has a stronger tendency to hallucinate than GPT4V. Researchers can use JourneyBench to diagnose where models excel and where they struggle to better target their research efforts. JourneyBench thus has the potential to significantly improve the accuracy and reliability of AI systems used in various applications with larger societal benefits, such as medical imaging, autonomous vehicles, and assistive technologies for people with disabilities. JourneyBench will allow fairer and broader comparison of AI models by providing a standardized benchmark where models can be compared and improved across a number of axes which have applications in critical downstream applications. Enhanced accuracy in these domains can lead to better diagnostics, safer transportation, and more effective assistance, thus improving overall quality of life. We expect models that will be compared on JourneyBench to be deployed in many sectors, such as intelligent tutoring and question answering. Further, because the datasets provided by JourneyBench are highly diverse and feature AI generated content, we expect JourneyBench to play an important role in benchmarking performance of AI systems on generated data, which we expect will continue to grow across social media and the internet, as well as benchmarking performance on unusual situations. Due to its rich diversity of atypical situations and content, JourneyBench will help in creating more robust and less biased AI models which is critical for deployment of AI systems in real world applications.

Potential Negative Impacts. On the other hand, there are potential negative societal impacts associated with the use and development of JourneyBench. One major concern is the exacerbation of existing biases within AI systems. JourneyBench was harvested from data generated from human prompts on MidJourney with models trained on images harvested from the web. If the data used to train these models was not carefully curated to avoid reinforcing stereotypes or excluding certain groups, the resulting generated images can perpetuate or even amplify societal inequalities by reflecting those biases within the data. While JourneyBench was harvested by humans who inspected samples from MidJourney, it is possible that some of these inequities exist within the data, despite being manually chosen (e.g. overrepresentation of certain racial groups). This may lead to biased analysis of models, such as by overestimating their performance on images containing minorities. More broadly, JourneyBench will help facilitate the improvement of advanced AI systems which could lead to increased surveillance and erosion of privacy, as more sophisticated AI could be employed in ways that monitor and analyze individuals' behavior without their consent (e.g. automatically analyzing behaviors, predicting next steps, etc.). There is also the risk of job displacement in industries where these advanced AI systems are implemented, leading to economic and social challenges for affected workers. For example, JourneyBench reveals that many models continue to struggle on multi-image chain of thought reasoning. As these capabilities improve, workers whose roles involve such analysis are at risk of replacement. To address these issues, we intend to address potential negative impacts through transparency, explicit ethical consideration statements, and policies that ensure AI development aligns with societal values and needs. For example, we will make clear that analysis on JourneyBench may reflect underlying biases.

M Limitations

One primary limitation of JourneyBench is the inherent difficulty in curating truly unbiased and representative imaginary images. While JourneyBench aims to test models in unusual and imaginative scenarios, the selection of these scenarios might still reflect certain biases or gaps. For instance, the types of imaginary images and tasks chosen might not cover all possible edge cases or cultural contexts, potentially limiting the generalizability of the benchmark's findings. Additionally, the reliance on generated images, although mitigating copyright issues and enabling diverse content, may introduce artifacts or inconsistencies that are not present in real-world images, potentially skewing the evaluation results. Because all generated images were harvested from the Midjourney website, generated images may contain biases or artifacts present in the AI models available at this time. For example, many image generators rely on conditioning from CLIP. If certain visual content is not well captured by CLIP's conditioning, it may not appear in the generated output. Further, as generative models advance in the coming years, new classes of models and conditioning may emerge. Those

25

models may contain a different set of artifacts or biases than present in JourneyBench, so performance on JourneyBench may not necessarily translate to those. In particular, some models we evaluate rely on CLIP's conditioning. If CLIP is also used in image generation, this may introduce a bias towards models relying on these encoders.

Another limitation of JourneyBench is that the tasks within it are designed to be extremely challenging and require complex, fine-grained visual reasoning. This focus on fine-grained details and unusual scenarios may not fully capture the broad utility of these models in more conventional applications, potentially underrepresenting their strengths in real-world tasks. Other limitations include the focus on English-language understanding (in all captions and question answering tasks), as opposed to other languages. This may further bias JourneyBench towards certain types of content found in English-speaking countries. Lastly, JourneyBench does not include any generated video understanding tasks. Prompt-based generated videos can be expected to proliferate in the coming years, with impressive results showcased by OpenAI's SORA. JourneyBench currently focuses on image understanding (including multi-image understanding), but does not currently address temporal understanding in generated videos.

N   Personally Identifiable Information and Offensive Content

The JourneyBench dataset is constructed with a strict focus on ethical standards and user safety. It does not contain any personally identifiable information (PII) or sensitive data related to individuals. All images in the dataset are generated and publicly posted for sharing through the Midjourney platform under the community rules, ensuring that no PII is included. Additionally, the dataset has been curated to exclude any content that might be considered offensive, insulting, threatening, or anxiety-inducing. The images underwent a multi-layered filtering process, initially by the Midjourney platform and subsequently through multiple rounds of human annotation, to ensure appropriateness and non-distressful content. This rigorous curation process guarantees that the JourneyBench dataset is suitable for a broad audience and aligns with ethical guidelines for public research and academic use. Therefore, individual consent for data collection is not applicable. The annotations were created by human annotators specifically for research purposes, ensuring that all data within JourneyBench is ethically sourced and suitable for academic and non-commercial research.

N.1   Digital Object Identifier

We have requested a DOI for JourneyBench on `https://registry.identifiers.org/` and await their approval.

N.2   HaloQuest Data

JourneyBench includes a task, VQA with hallucination triggers, which is derived from a previous work titled "HaloQuest: A Visual Hallucination Dataset for Advancing Multimodal Reasoning." HaloQuest is currently under review and planned for release soon. The authors of this work are responsible for both the HaloQuest and JourneyBench data. There are no ethical issues in HaloQuest beyond those already addressed in JourneyBench.

O   Future Maintenance Plan

The JourneyBench dataset will undergo regular updates and maintenance to ensure its continued relevance and accuracy in evaluating multimodal models. The research team at Columbia University, UCLA, and Virginia Tech will be responsible for these updates, which will include correcting labeling errors, adding new instances, and removing outdated or erroneous data. Updates will be communicated to users through the official GitHub repository at `https://github.com/JourneyBench/JourneyBench`, the project website at `https://journeybench.github.io/`, and a mailing list for subscribed users. The team aims to review and update the dataset at least quarterly or more frequently as needed based on feedback and the identification of new challenges in the field. The maintenance would continue for at least five years after the paper's acceptance. Additionally, a leaderboard will be developed to track and document future works and their model performance

26

using the JourneyBench dataset, fostering a collaborative environment for ongoing research and improvement.

We plan to share the dataset on Hugging Face and host a workshop focusing on a competition via JourneyBench at the upcoming CVPR conference. These initiatives will broaden access to the dataset and encourage active participation and collaboration within the research community.

P   Terms of Usage for JourneyBench Dataset

P.1   Ownership and Responsibility

The JourneyBench dataset contains images obtained from the Internet, including those generated by Midjourney, which are not the property of Columbia University, UCLA, or Virginia Tech. These institutions are not responsible for the content or meaning of these images.

The authors state that to the best of their knowledge, information, and belief they have obtained all content in JourneyBench from sources such as Midjourney which allow for the intended use and redistribution in JourneyBench. The authors assume full responsibility for violation of any rights from content in JourneyBench and will immediately move to rectify any such violation should such violation be brought to the authors' attention. All data was harvested consistent with the Terms of Use of Midjourney and other platforms used by the authors to create and assemble JourneyBench.

Fair use notice. The authors acknowledge that in the United States, copyright of generative content remains an issue in flux. Should any generated content within JourneyBench ever be held to fall under copyright under current US law, JourneyBench can still be distributed under fair use. Specifically, we make JourneyBench available in an effort to advance understanding of technological, scientific, and cultural issues. We believe this constitutes a 'fair use' of any such copyrighted material as provided for in Section 107 of the US Copyright Law. In accordance with Title 17 U.S.C. Section 107, the material in JourneyBench is distributed without profit to those who have expressed a prior interest in receiving the included information for non-commercial research and educational purposes. For more information on fair use please click here. If you wish to use copyrighted material in JourneyBench for purposes of your own that go beyond non-commercial research and academic purposes, you must obtain permission directly from the copyright owner should one exist.

P.2   Non-commercial Research

The JourneyBench dataset is **ONLY** available for non-commercial research purposes. Any use of the dataset for commercial purposes is strictly prohibited.

P.3   Competitive Research

You may not use the JourneyBench dataset for competitive research against Midjourney or any other image generation platforms.

P.4   Restrictions on Usage

- You agree not to reproduce, duplicate, copy, sell, trade, resell, or exploit any portion of the images or derived data for commercial purposes.
- You agree not to further copy, publish, or distribute any portion of the JourneyBench dataset.
- Except for internal use at a single site within the same organization, making copies of the dataset is prohibited.

P.5   Interpretation and Revision

The research team at Columbia University, UCLA, and Virginia Tech reserves the right to interpret and revise these terms.

27

P.6 Removal of Product

If you do not wish to have your product included in the JourneyBench dataset, please contact us at journeybench.contact@gmail.com to have it removed.

By using the JourneyBench dataset, you agree to comply with these terms of usage. Any violation of these terms may result in the termination of your access to the dataset and could lead to legal action.

P.7 Licensing

The JourneyBench dataset is distributed under a custom license that includes the following terms based on the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license, with additional restrictions:

- **Attribution**: You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
- **NonCommercial**: You may not use the material for commercial purposes.
- **NoDerivatives**: If you remix, transform, or build upon the material, you may not distribute the modified material.
- **Additional Restrictions**: The dataset may not be used for competitive research against Midjourney or any other image generation platforms. You also agree not to further copy, publish, or distribute any portion of the dataset beyond internal use at a single site within the same organization.

For more details, visit `https://creativecommons.org/licenses/by-nc-nd/4.0/`.

By incorporating these terms, the JourneyBench dataset can be distributed in a manner that respects the privacy and usage policies of the original sources, while also ensuring it is used appropriately within the research community.

References

[1] Llama-3. `https://ai.meta.com/blog/meta-llama-3/`

[2] Midjourney. `https://midjourney.com/`

[3] Bitton-Guetta, N., Bitton, Y., Hessel, J., Schmidt, L., Elovici, Y., Stanovsky, G., Schwartz, R.: Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2616–2627 (2023)

[4] Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)

[5] Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Muyan, Z., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. arXiv preprint arXiv:2312.14238 (2023)

[6] Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2818–2829 (2023)

[7] Chun, S., Kim, W., Park, S., Chang, M., Oh, S.J.: Eccv caption: Correcting false negatives by collecting machine-and-human-verified image-caption associations for ms-coco. In: European Conference on Computer Vision. pp. 1–19. Springer (2022)

[8] Jiang, D., He, X., Zeng, H., Wei, C., Ku, M., Liu, Q., Chen, W.: Mantis: Interleaved multi-image instruction tuning. arXiv preprint arXiv:2405.01483 (2024)

28

[9] Laurençon, H., Tronchon, L., Cord, M., Sanh, V.: What matters when building vision-language models? arXiv preprint arXiv:2405.02246 (2024)

[10] Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)

[11] Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems **34**, 9694–9705 (2021)

[12] Lin, J., Yin, H., Ping, W., Lu, Y., Molchanov, P., Tao, A., Mao, H., Kautz, J., Shoeybi, M., Han, S.: Vila: On pre-training for visual language models. arXiv preprint arXiv:2312.07533 (2023)

[13] Liu, H., Singh, P.: Conceptnet—a practical commonsense reasoning tool-kit. BT technology journal **22**(4), 211–226 (2004)

[14] Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. Advances in Neural Information Processing Systems **35**, 2507–2521 (2022)

[15] Pan, J., Sun, K., Ge, Y., Li, H., Duan, H., Wu, X., Zhang, R., Zhou, A., Qin, Z., Wang, Y., Dai, J., Qiao, Y., Li, H.: Journeydb: A benchmark for generative image understanding (2023)

[16] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)

[17] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2022)

[18] Wang, H., Lin, G., Hoi, S.C.H., Miao, C.: Paired cross-modal data augmentation for fine-grained image-to-text retrieval (2022)

[19] Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., et al.: Image as a foreign language: Beit pretraining for all vision and vision-language tasks. arXiv preprint arXiv:2208.10442 (2022)

[20] Wang, X., Zhou, Y., Liu, X., Lu, H., Xu, Y., He, F., Yoon, J., Lu, T., Bertasius, G., Bansal, M., Yao, H., Huang, F.: Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences (2024)

[21] Wang, Z., Bingham, G., Yu, A., Le, Q., Luong, T., Ghiasi, G.: Haloquest: A visual hallucination dataset for advancing multimodal reasoning (2024)

[22] Yeh, M.H., Chen, V., Haung, T.H.K., Ku, L.W.: Multi-vqg: Generating engaging questions for multiple images (2022)

[23] Zeng, Y., Zhang, X., Li, H., Wang, J., Zhang, J., Zhou, W.: X 2-vlm: All-in-one pre-trained model for vision-language tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)

[24] Zou, X., Wu, C., Cheng, L., Wang, Z.: Tokenflow: Rethinking fine-grained cross-modal alignment in vision-language retrieval (2022)

29