

SMIR: A SYNTHETIC DATA PIPELINE TO IMPROVE MULTI-IMAGE REASONING

Anonymous authors

Paper under double-blind review

ABSTRACT

Vision-Language Models (VLMs) have demonstrated strong performance in single-image understanding, supported by many high-quality instruction datasets. However, multi-image reasoning tasks remain under-explored in the open-source community due to two major issues: (1) scaling up datasets with multiple correlated images and complex reasoning instructions is resource-intensive and difficult to maintain quality and (2) there is a shortage of robust multi-image evaluation benchmarks. To address these issues, we introduce SMIR, an efficient synthetic data-generation pipeline for multi-image reasoning, and a high-quality SMIR dataset generated using this pipeline. Our pipeline efficiently extracts highly correlated images using multimodal embeddings, combining visual and descriptive information and leverages open-source LLMs to generate quality instructions, offering a cost-effective alternative to expensive closed-source solutions. Additionally, we present SMIR-BENCH, a novel multi-image reasoning evaluation benchmark comprising 100 diverse examples across 7 complex multi-image reasoning tasks. Unlike existing benchmarks, SMIR-BENCH is multi-turn and allows for free-form responses, providing a more comprehensive evaluation of model expressiveness and reasoning capability. We demonstrate the effectiveness of SMIR dataset by fine-tuning several open-source VLMs and evaluating their performance on SMIR-BENCH. Our results show that models trained on our dataset outperform baseline models in multi-image reasoning tasks. Furthermore, we observe enhanced model expressiveness and more nuanced reasoning in free-form responses, highlighting the value of our approach for advancing open-source VLM research.¹

1 INTRODUCTION

Vision-Language Models (VLMs) have shown impressive capabilities in tasks involving single images, particularly open-source models that have benefited from high-quality instruction datasets (Laurençon et al., 2024b; Zhang et al., 2023; Xu et al., 2022). However, when it comes to multi-image tasks, such as comparing or analyzing relationships between multiple images, the performance of open-source VLMs (Liu et al., 2024b;a; Li et al., 2024a; Awadalla et al., 2023; Yao et al., 2024; Wang et al., 2023b) lags significantly behind their closed-source counterparts in GPT-4 (Achiam et al., 2023), Claude 3.5 Sonnet (Anthropic, 2024a), Claude 3 (Anthropic, 2024b), and Gemini 1.5 (Reid et al., 2024). One of the crucial problems is that constructing large-scale, complicated multi-image reasoning datasets and evaluation benchmarks is challenging.

First, collecting and curating large-scale multiple images with high correlations is hard. Identifying correlated semantic information or entities across images requires large-scale images and sophisticated algorithms. Thus, most existing multi-image instruction tuning datasets do not have highly correlated images. For example, MANTIS (Jiang et al., 2024) often includes unrelated images within the same multi-image reasoning question, potentially undermining the complexity of the task. MMDU-45K (Liu et al., 2024e) attempts to address this by clustering image captions. MMInstruct (Liu et al., 2024d), on the other hand, only considers one image at a time, falling short

¹Upon acceptance, we will open-source the synthetic data generation pipeline, our dataset, and evaluation benchmark.

of true multi-image reasoning. These shortcomings highlight the difficulties and needs for datasets featuring related images within multi-image scenarios.

Second, scaling up the number of highly correlated images presents significant challenges. Existing datasets such as MANTIS, MMDU-45K, and MultiInstruct require extensive human curation and annotation, resulting in a labor-intensive and time-consuming process. To minimize human effort, researchers have leveraged GPT-4 family models (Peng et al., 2023; Wang et al., 2023a) to generate synthetic datasets—including MMInstruct, Multimodal ArXiv (Li et al., 2024b), MIMIC-IT (Li et al., 2023a), StableLLaVA-Instruct (Li et al., 2023c), and SVIT-Instruct (Zhao et al., 2023). However, this method proves expensive and difficult to scale effectively.

Third, evaluating multi-image reasoning is complicated. Given the increased complexity of multi-image reasoning tasks, using multi-turn free-response evaluation instead of the multiple-choice format employed by previous benchmarks such as (Fu et al., 2024; Ying et al., 2024; Wang et al., 2024a; Yue et al., 2024; Singh et al., 2019; Hudson & Manning, 2019; Antol et al., 2015). Free-response evaluations are more challenging, requiring models to articulate their thought processes, providing insight into their reasoning abilities, and allowing for a more nuanced assessment of their capabilities.

To address these challenges, we propose a synthetic data generation pipeline, SMIR, for multi-image reasoning and a human-annotated evaluation benchmark for multi-image reasoning, SMIR-BENCH. SMIR aims to generate correlated and challenging multi-image reasoning questions, while SMIR-BENCH evaluates models on free-response, difficult multi-image scenarios.

To summarize, we address these issues with **our contributions**:

- Two novel sampling algorithms: Cluster Sampling for data quality robustness and Graph Iteration Sampling for diversity. Both use multimodal embeddings (combining image and caption) to group correlated images for challenging multi-image instruction tuning datasets.
- A scalable synthetic multimodal data generation pipeline utilizing open-source LLMs such as Meta Llama 3.1 70B Instruct Turbo (Dubey et al., 2024), eliminating the need for expensive closed-source models, reducing costs by up to 50 times (Kirkovska, 2024) and speeds up to by 10 times (Kirkovska, 2024), while significantly minimizing human annotation efforts.
- A new multimodal evaluation benchmark with free-form responses, assessing both final answers and reasoning processes in complex multi-image tasks. Using GPT-4-Turbo and other open-source models as reference, we see up to 11% improvement with the SMIR dataset.

2 RELATED WORKS

Vision Language Models We focus on instruction tuning Vision-Language Models (VLMs) that utilize a pretrained Large Language Model (LLM) backbone because this approach is cost-effective and more accessible for the open-source community. Since the backbone responsible for language understanding is already trained, the overall training process becomes simpler and requires fewer resources. Our primary task involves aligning the vision encoder—typically architectures like Vision Transformer (ViT) (Dosovitskiy, 2020), SigLIP (Zhai et al., 2023), or CLIP (Radford et al., 2021)—with the LLM backbone. This alignment is facilitated through linear layers that connect the vision encoder to the backbone, enabling the integration of visual and textual information. For instance, BLIP-2 (Li et al., 2023b) uses OPT (Zhang et al., 2022) and FLAN-T5 (Chung et al., 2022) as backbones, MiniGPT-4 (Zhu et al., 2023) utilizes Vicuna (Chiang et al., 2023), and Qwen-2-VL (Wang et al., 2024b) employs Qwen-2-1.5B (Yang et al., 2024) as the language backbone. In this paper, we focus on creating a high-quality multi-image reasoning dataset for instruction tuning instead of large-scale interleaved pretraining datasets like OBELICS (Laurençon et al., 2024a), MINT-1T (Awadalla et al., 2024), and LAION-5B (Schuhmann et al., 2022).

Multi-Image Reasoning Data Recent advancements in multi-image reasoning instruction tuning datasets include MANTIS (Jiang et al., 2024) and MMDU-45K (Liu et al., 2024e), both aiming to improve reasoning capabilities in VLMs. However, these datasets have limitations in their approaches. MANTIS randomly concatenates single image pairs from LLaVA-665k (Liu et al., 2024a),

108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123

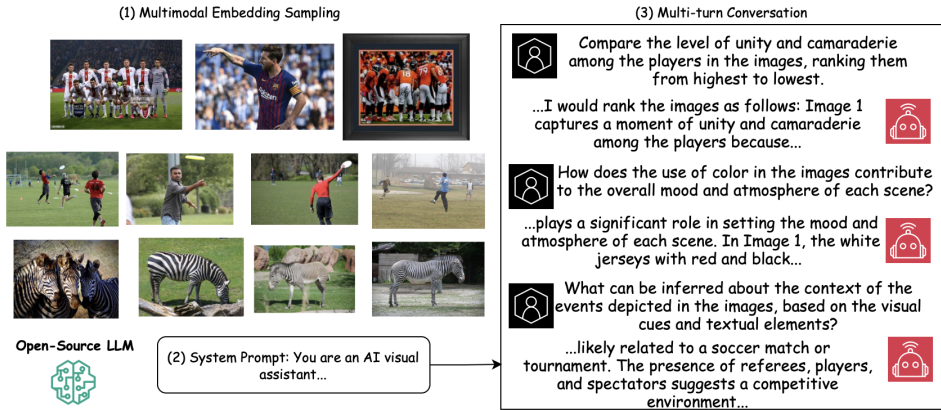


Figure 1: Our end-to-end pipeline: from sampling and LLM prompting to conversation generation. The example conversation is based on a sports scenario, demonstrating the pipeline’s ability to generate contextually relevant multi-turn dialogues.

124
125
126
127
128
129
130
131
132
133
134
135
136
137

which often results in uncorrelated images within multi-image scenarios, potentially undermining the complexity of reasoning tasks. MMDU-45K attempts to address this issue by utilizing sentence transformers (Reimers, 2019) with description text and clustering techniques to group related images, but does not consider visual components. The dataset is then further enhanced, assisted by GPT-4 to generate comprehensive answers for the grouped images. Building upon these efforts, SMIR introduces a novel approach that leverages both vital visual and caption information to ensure highly correlated images within multi-image sets with the use of open-source LLMs. These scalable methods leads to the generation of more challenging questions that require deeper analysis and understanding of visual relationships, pushing the boundaries of multi-image reasoning capabilities in VLMs.

Datasets	Multimodal Embedding	Correlated Images	Human-Annotation	Open-Source LLM
Mantis	No	No	Yes	No
MMDU	No	Yes	Yes	No
SMiR	Yes	Yes	No	Yes

138
139

Table 1: Comparison of datasets highlighting key characteristics and methodologies.

140
141
142
143
144
145
146
147
148
149
150
151
152
153
154

Multi-Image Reasoning Benchmarks Recent VLM benchmarks (Chiang et al., 2024; Lin et al., 2024; Liu et al., 2024c) have made strides by incorporating free-response evaluations, marking a significant improvement over traditional multiple-choice formats. However, these benchmarks still lack a comprehensive approach that combines automatic, multi-turn, and pairwise evaluation capabilities. Our benchmark addresses this gap, drawing inspiration from Auto-Hard-Auto v0.1 (Li et al., 2024c). We have adapted and expanded this framework to enable robust multimodal evaluation, providing a more holistic assessment of VLM performance across complex, multi-image reasoning tasks. This approach allows for a deeper analysis of both the final answers and the underlying reasoning processes employed by VLMs in real-world SMIR-BENCH scenarios.

155
156
157

3 SMIR: SYNTHETIC MULTI-IMAGE REASONING DATA PIPELINE

158
159
160
161

To generate complicated multi-image reasoning synthetic data efficiently, we introduce the SMIR pipeline. Given a large-scale of image-caption dataset D with N pairs of image-captions in D as $(I_i, C_i)_{i=1}^N \in D$. SMIR constructs a multimodal embedding E_i for each pair of (I_i, C_i) . Then, we apply grouping algorithms to find the correlations between multimodal pairs based on the embeddings. Finally, open-sourced LLMs are prompted to generate complicated question-answering

162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215

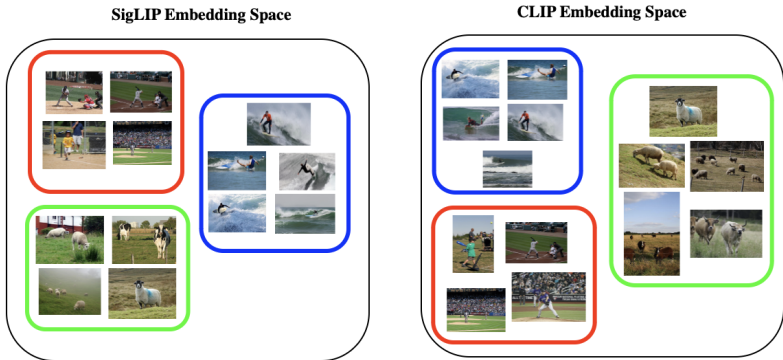


Figure 2: Cluster Matching from Different Embedding Spaces. Images are sampled only from the union of clusters with the same color.

pairs for on multiple pairs sampled based on correlations. In this section, we first introduce how we construct the multimodal embeddings, then we present the grouping algorithms, and finally we show how we generate the synthetic data samples via open-source LLM.

3.1 MULTIMODAL EMBEDDING CONSTRUCTION

To identify correlated images effectively, we developed a method that incorporates both visual and textual information from image-caption pairs. Our approach utilizes SigLIP and CLIP image embeddings alongside corresponding caption embeddings. We formulated a multimodal embedding by combining these components with a small constant, c , as follows:

$$E_{multimodal} = E_{image} + c \cdot E_{caption} \tag{1}$$

where $E_{multimodal}$ is the multimodal embedding, E_{image} is the image embedding, and $E_{caption}$ is the caption embedding. For the ShareGPT4V (Chen et al., 2023), a $c = 0.2$ worked well, but this parameter may vary depending on the quality of individual image-caption pairs in other data sources.

Importantly, **relying solely on either image or caption information would limit our ability to concurrently consider both visual and textual contexts, which is crucial for establishing a comprehensive understanding of the images.** This multimodal approach enables us to capture the nuanced relationships between visual content and its associated descriptive text, thereby enhancing our capacity to identify and group correlated images effectively.

Following the generation of multimodal embeddings, we employed UMAP (McInnes et al., 2018) to reduce the dimensionality of the vectors. This technique allowed us to project the high-dimensional embeddings into a lower-dimensional space, facilitating more efficient analysis and visualization of the data while preserving its essential structure.

3.2 GROUPING IMAGES

We present two novel algorithms designed to group correlated images prior to leveraging an open-source Large Language Model (LLM) for synthetic data generation in multi-image reasoning tasks. The emphasis on correlated images is crucial, as it facilitates challenging multi-image reasoning scenarios. These scenarios require the model to identify intricate relationships and differentiate between visually similar scenes, thus enhancing the complexity and realism of the reasoning process.

Clustering We employed HDBSCAN (Malzer & Baum, 2020), a density-based clustering algorithm, to group the SigLIP and CLIP multimodal embeddings into coherent clusters. To establish meaningful relationships between the two embedding spaces, we developed a greedy algorithm that

216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269

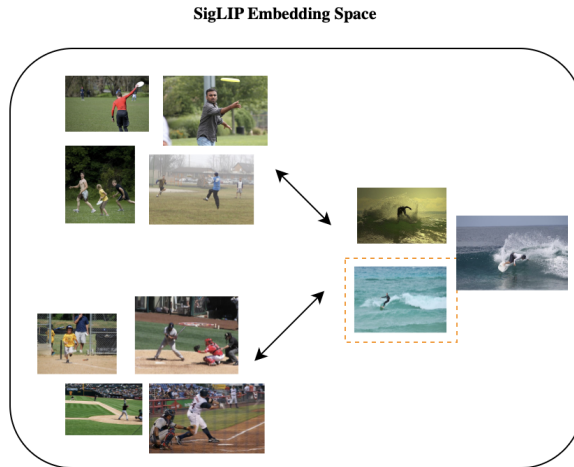


Figure 3: Sampling based on embedding distance for increased diversity from initial image (orange). All images are considered within a single question.

matched SigLIP and CLIP clusters in a one-to-one fashion. This matching process ensured that each cluster from one embedding model corresponded to a semantically similar cluster from the other. The detailed steps of this Greedy Cluster Matching Algorithm in Figure 2 are presented in Algorithm 1 (Appendix A.1). Once clusters are matched, images are sampled from the cluster union until the desired number of images is selected.

Vector Space Sampling We developed an iterative sampling method to select diverse yet related images within each embedding space. The process begins by randomly selecting an initial image, then iteratively sampling subsequent images based on their distance from previously selected points in the embedding space. This approach continues until the desired number of images is reached, ensuring a balance between diversity and semantic coherence in the selected image set. The detailed steps of this Random Sample Iteration algorithm in Figure 3 are presented in Algorithm 2 (Appendix A.2).

By assembling related images before prompting the LLM, we create a more coherent and contextually rich input, enabling the model to generate more nuanced and relevant synthetic data for multi-image reasoning tasks.

3.3 GENERATE SYNTHETIC DATA

Once grouped image-caption pairs are sampled, we take the caption embeddings and incorporate them into a system prompt for an open-source LLM, such as Meta Llama 3.1 70B Turbo, up to 50 times cheaper and 10 times faster compared to GPT-4 (Kirkovska, 2024). This process generates complex multi-turn conversations between User and Assistant as seen in Table 2 and questions tailored to the selected images, as shown in Figure 6 and Figure 7 (Appendix B).

4 DISCUSSION

Our approach involves several design choices, each with its own trade-offs. In this section, we discuss the decisions behind sampling algorithms, prompts, and data sources.

4.1 SAMPLING ALGORITHMS

Cluster-based algorithms demonstrate high efficacy in producing quality image-caption pairs, leveraging both SIGLIP and CLIP embeddings to confirm spatial relationships and associated semantic meanings. This dual-embedding validation ensures robust data quality, as images matched within

Table 2: SMIR Dataset Statistics

Metric	Value
Number of Samples	160,000
Maximum Number of Turns	24
Minimum Number of Turns	2
Average Number of Turns	9.65
Average Number of Images	4.65
Average User Tokens	25.51
Average Assistant Tokens	124.32
Open-Source LLM	Meta Llama 3.1 70B Turbo

clusters are corroborated by two independent embedding models. However, this approach has a notable limitation: it can lead to overly specialized image subjects. This specialization occurs because sampling is confined to a single matched cluster, which restricts the diversity of selected images by excluding images from other clusters. A detailed example is presented in Figure 6 (Appendix B.1).

Vector sampling emerged as the preferred method due to its capacity to generate more generalized image subjects and foster diverse question generation when coupled with a system prompt. This approach allows for a wider range of image combinations, transcending the boundaries of individual clusters. Consequently, it facilitates the creation of more varied and cognitively demanding reasoning tasks. The flexibility of vector sampling in drawing from a broader semantic space contributes to a richer, more diverse dataset, potentially enhancing the complexity and applicability of subsequent machine learning tasks. A detailed example is presented in Figure 7 (Appendix B.2).

4.2 PROMPTS

In our data generation approach, we focused on creating two distinct types of questions: shorter, quick visual questions often involving OCR tasks, and longer reasoning questions that require in-depth analysis. Drawing inspiration from CoT (Wei et al., 2022), we designed prompts to generate multi-turn conversations, enhancing the complexity and depth of interactions. This dual approach necessitated the development of separate, tailored prompts for each question type, allowing us to effectively capture both complex reasoning scenarios and straightforward visual comprehension tasks. More details about the short prompt in Figure 8 (Appendix C.1 and long prompts Figure 9 (Appendix C.2).

4.3 DATA SOURCE

Our study leveraged the ShareGPT4V (Chen et al., 2023) dataset as the primary source for generating synthetic examples. This comprehensive dataset comprises of better image-caption pairs derived from LLaVA-Instruct (Liu et al., 2024a) and COCO (Lin et al., 2014). To maintain diversity from the data source, we synthetically generated 5,000 data points from each 20,000-image batch, resulting in a total of 160,000 synthetic examples. SMIR pipeline can also be applied easily to other data resources in the future.

5 MULTI-IMAGE BENCHMARK

SMIR-BENCH extends the Auto-Hard-Auto v0.1 (Li et al., 2024c) framework to the multimodal domain. It employs a judge model for pairwise comparison against a baseline model, evaluating responses on helpfulness, relevance, and conciseness. This approach enables a multi-turn, automatic, and challenging evaluation process.

5.1 MOTIVATIONS

We were motivated to create questions that more challenge VLMs to reason over multiple related images, analyze relationships, and derive meaning from series of images. We developed a multi-turn benchmark of 100 examples across seven diverse topics. This benchmark challenges models

to analyze relationships, derive meaning from image series, and provide hard explanations for complex visual tasks. Curated collaboratively by a human annotator and GPT-4, it uses images from the internet and Shot2Story (Han et al., 2023), compelling VLMs to demonstrate advanced visual reasoning capabilities beyond answering multiple choice.

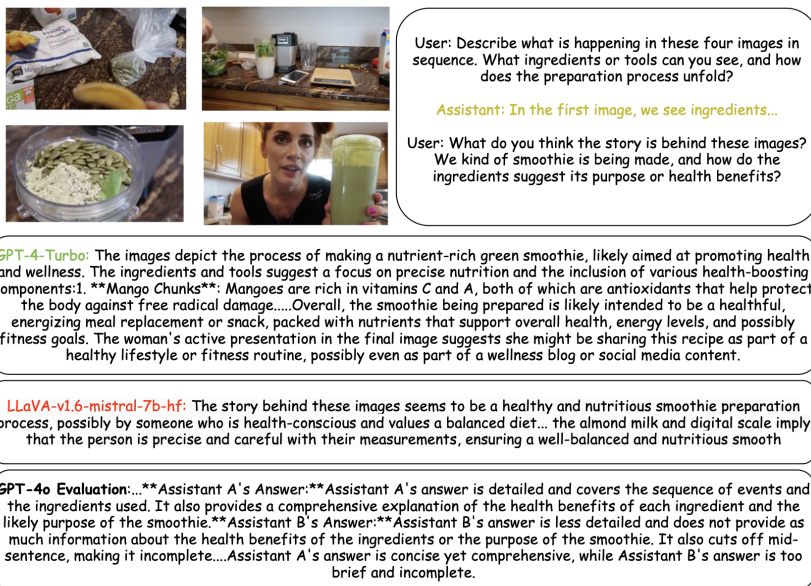


Figure 4: Evaluation Benchmark Using GPT-4o as Judge

5.2 EXPERIMENTS

We fine-tuned Mantis-8B-siglip-llama3-pretrained and idefics-8b on the SMiR dataset, using substantially less data (160K samples compared to the original 721K). These fine-tuned models were then evaluated on SMiR-BENCH. Remarkably, they outperformed both Mantis-8B-siglip-llama3 and Mantis-8B-Idefics2, despite the reduced training data. This improvement was evident when benchmarking against the closed-source GPT-4-Turbo, as well as against Mantis-8B-siglip-llama3 and Mantis-8B-Idefics2 themselves. These results demonstrate the effectiveness of our fine-tuning approach on the SMiR dataset, achieving superior performance with performance gains.

Table 3: Model Scores with GPT-4-Turbo Baseline

Model Name	Score	Δ	95% CI	Average Tokens
GPT-4o	68.1		(-5.3, 6.4)	442
Claude-3.5-Sonnet-20240620	54.7		(-4.6, 7.0)	374
GPT-4-Turbo	50.0		(0.0, 0.0)	377
Gemini-1.5-Pro	38.7		(-7.0, 5.6)	349
Claude-3-Opus-20240229	31.0		(-5.7, 7.6)	338
Mantis-8B-siglip-llama3-pretrained + SMiR-160k	9.5	+3.4%	(-3.1, 3.1)	180
Mantis-8B-siglip-llama3	6.1		(-2.5, 2.6)	170
Idefics2-8B + SMiR-50k	6.0	+.6%	(-2.3, 2.9)	173
Mantis-8B-Idefics2	5.4		(-2.4, 2.3)	195
Idefics2-8B	4.6		(-2.7, 2.0)	122
LLaVA-v1.6-mistral-7b-hf	2.5		(-1.2, 1.6)	361
Mantis-8B-siglip-llama3-pretrained	2.2		(-1.6, 1.8)	198

Table 4: Model Scores with Mantis-8B-siglip-llama3 baseline

Model Name	Score	Δ	95% CI	Average Tokens
Claude-3-Opus-20240229	96.9		(-1.9, 2.0)	338
Claude-3-5-Sonnet-20240620	96.3		(-2.3, 1.4)	374
GPT-4-Turbo	95.0		(-2.1, 2.5)	377
Gemini-1.5-Pro	94.2		(-2.7, 2.1)	349
GPT-4o	91.8		(-3.4, 3.1)	442
Mantis-8B-siglip-llama3-pretrained + SMiR-160k	57.0	+7.0%	(-7.9, 6.9)	180
Mantis-8B-siglip-llama3	50.0		(0.0, 0.0)	170
LLaVA-v1.6-Mistral-7B-HF	18.9		(-4.0, 4.4)	361
Mantis-8B-siglip-llama3-pretraind	11.7		(-4.2, 5.2)	198

Table 5: Model Scores with Mantis-8B-Idefics2 Baseline

Model Name	Score	Δ	95% CI	Average Tokens
Claude-3-Opus-20240229	97.6		(-2.0, 1.4)	338
Gemini-1.5-Pro	97.2		(-2.2, 1.4)	349
Claude-3-5-Sonnet-20240620	96.8		(-1.7, 1.5)	374
GPT-4-Turbo	94.3		(-2.6, 2.4)	377
GPT-4o	93.0		(-3.2, 2.8)	442
Idefics2-8B + SMiR-50k	61.0	+11.0%	(-7.4, 7.7)	173
Mantis-8B-Idefics2	50.0		(0.0, 0.0)	195
Idefics2-8B	31.2		(-5.5, 5.9)	122
LLaAV-v1.6-mistral-7b-hf	20.1		(-5.0, 3.9)	361

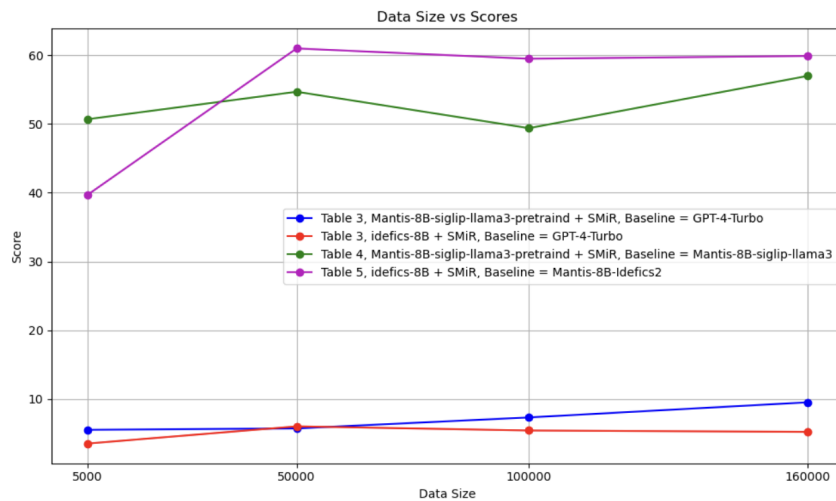


Figure 5: SMiR Dataset Size vs. Benchmark Score

6 CONCLUSION

This paper introduces a synthetic data pipeline designed to enhance multi-image reasoning capabilities on open-source VLMs. By leveraging multimodal embeddings and grouping algorithms, the pipeline generates high-quality synthetic multi-image reasoning instruction tuning data. The approach yields up to 11% improvement on SMiR-BENCH for popular open-source models, demonstrating the significant potential of synthetic data in advancing open-source VLM models.

Limitations Our methods have several limitations. Random sampling with iteration is time-intensive due to the need for recalculating distance embeddings for each new image sampled. Further investigation is needed to determine the scalability of our synthetic data. Future research should focus on developing more time-efficient algorithms and optimizing data mixtures.

Broader Impact This paper introduces a method for generating high-quality, cost-effective data for VLMs, addressing the growing challenge of data scarcity. By advancing these open-source techniques, we contribute to narrowing the performance gap between open and closed-source models, promoting more accessible and powerful multimodal AI.

7 REPRODUCIBILITY STATEMENT

The data sources are available on ShareGPT4V (Chen et al., 2023). Grouping algorithm codes can be found in (Appendix A, and prompts are provided in (Appendix C). All exact codes will be released and open-source.

REFERENCES

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Anthropic. Claude 3.5 sonnet model card addendum, 2024a. URL <https://www.anthropic.com/model-card/claude-3-5-sonnet-addendum>. Accessed: 2024-09-30.
- Anthropic. Claude 3 model card, 2024b. URL <https://www.anthropic.com/model-card/claude-3>. Accessed: 2024-09-30.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.
- Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.
- Anas Awadalla, Le Xue, Oscar Lo, Manli Shu, Hannah Lee, Etash Kumar Guha, Matt Jordan, Sheng Shen, Mohamed Awadalla, Silvio Savarese, et al. Mint-1t: Scaling open-source multimodal data by 10x: A multimodal dataset with one trillion tokens. *arXiv preprint arXiv:2406.11271*, 2024.
- Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL <https://lmsys.org/blog/2023-03-30-vicuna/>.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

- 486 Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale.
487 *arXiv preprint arXiv:2010.11929*, 2020.
488
- 489 Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha
490 Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models.
491 *arXiv preprint arXiv:2407.21783*, 2024.
- 492 Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A
493 Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but
494 not perceive. *arXiv preprint arXiv:2404.12390*, 2024.
- 495 Mingfei Han, Linjie Yang, Xiaojun Chang, and Heng Wang. Shot2story20k: A new benchmark for
496 comprehensive understanding of multi-shot videos. *arXiv preprint arXiv:2311.17043*, 2023.
497
- 498 Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning
499 and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer
500 vision and pattern recognition*, pp. 6700–6709, 2019.
- 501 Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhua Chen. Mantis:
502 Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024.
503
- 504 Anita Kirkovska. Llama 3 70b vs gpt-4: Comparison analysis. *Vellum Blog*, 2024. URL <https://www.vellum.ai/blog-post-categories/model-comparison>. Accessed: 2024-
505 09-30.
506
- 507 Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov,
508 Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open
509 web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information
510 Processing Systems*, 36, 2024a.
- 511 Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building
512 vision-language models? *arXiv preprint arXiv:2405.02246*, 2024b.
513
- 514 Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan
515 Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint
516 arXiv:2306.05425*, 2023a.
- 517 Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei
518 Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint
519 arXiv:2408.03326*, 2024a.
- 520 Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image
521 pre-training with frozen image encoders and large language models. In *International conference
522 on machine learning*, pp. 19730–19742. PMLR, 2023b.
- 523 Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multi-
524 modal arxiv: A dataset for improving scientific comprehension of large vision-language models.
525 *arXiv preprint arXiv:2403.00231*, 2024b.
526
- 527 Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gon-
528 zalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and
529 benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*, 2024c.
- 530 Yanda Li, Chi Zhang, Gang Yu, Zhibin Wang, Bin Fu, Guosheng Lin, Chunhua Shen, Ling Chen,
531 and Yunchao Wei. Stablellava: Enhanced visual instruction tuning with synthesized image-
532 dialogue data. *arXiv preprint arXiv:2308.10253*, 2023c.
533
- 534 Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina
535 Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. Wildbench: Benchmarking llms with
536 challenging tasks from real users in the wild. *arXiv preprint arXiv:2406.04770*, 2024.
- 537 Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr
538 Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer
539 Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014,
Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

- 540 Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction
541 tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*,
542 pp. 26296–26306, 2024a.
- 543 Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances*
544 *in neural information processing systems*, 36, 2024b.
- 546 Shuo Liu, Kaining Ying, Hao Zhang, Yue Yang, Yuqi Lin, Tianle Zhang, Chuanhao Li, Yu Qiao,
547 Ping Luo, Wenqi Shao, et al. Convbench: A multi-turn conversation evaluation benchmark
548 with hierarchical capability for large vision-language models. *arXiv preprint arXiv:2403.20194*,
549 2024c.
- 550 Yangzhou Liu, Yue Cao, Zhangwei Gao, Weiyun Wang, Zhe Chen, Wenhai Wang, Hao Tian, Lewei
551 Lu, Xizhou Zhu, Tong Lu, et al. Mminstruct: A high-quality multi-modal instruction tuning
552 dataset with extensive diversity. *arXiv preprint arXiv:2407.15838*, 2024d.
- 554 Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong,
555 Yu Qiao, Dahua Lin, et al. Mmdu: A multi-turn multi-image dialog understanding benchmark
556 and instruction-tuning dataset for lvlms. *arXiv preprint arXiv:2406.11833*, 2024e.
- 557 Claudia Malzer and Marcus Baum. A hybrid approach to hierarchical density-based cluster selec-
558 tion. In *2020 IEEE international conference on multisensor fusion and integration for intelligent*
559 *systems (MFI)*, pp. 223–228. IEEE, 2020.
- 561 Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and
562 projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.
- 563 Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning
564 with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.
- 566 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal,
567 Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual
568 models from natural language supervision. In *International conference on machine learning*, pp.
569 8748–8763. PMLR, 2021.
- 570 Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-
571 baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gem-
572 ini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint*
573 *arXiv:2403.05530*, 2024.
- 575 N Reimers. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint*
576 *arXiv:1908.10084*, 2019.
- 577 Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi
578 Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An
579 open large-scale dataset for training next generation image-text models. *Advances in Neural*
580 *Information Processing Systems*, 35:25278–25294, 2022.
- 581 Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh,
582 and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF*
583 *conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.
- 585 Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma,
586 Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust
587 multi-image understanding. *arXiv preprint arXiv:2406.09411*, 2024a.
- 588 Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. To see is to
589 believe: Prompting gpt-4v for better visual instruction tuning. *arXiv preprint arXiv:2311.07574*,
590 2023a.
- 591 Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu,
592 Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the
593 world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

- 594 Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang,
595 Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv*
596 *preprint arXiv:2311.03079*, 2023b.
- 597 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny
598 Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in*
599 *neural information processing systems*, 35:24824–24837, 2022.
- 600 Zhiyang Xu, Ying Shen, and Lifu Huang. Multiinstruct: Improving multi-modal zero-shot learning
601 via instruction tuning. *arXiv preprint arXiv:2212.10773*, 2022.
- 602 An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li,
603 Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint*
604 *arXiv:2407.10671*, 2024.
- 605 Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li,
606 Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding
607 Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong
608 Sun. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint 2408.01800*, 2024.
- 609 Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang,
610 Yuqi Lin, Shuo Liu, et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating
611 large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*, 2024.
- 612 Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens,
613 Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multi-
614 modal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF*
615 *Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.
- 616 Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language
617 image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer*
618 *Vision*, pp. 11975–11986, 2023.
- 619 Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi
620 Hu, Tianwei Zhang, Fei Wu, et al. Instruction tuning for large language models: A survey. *arXiv*
621 *preprint arXiv:2308.10792*, 2023.
- 622 Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christo-
623 pher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer
624 language models. *arXiv preprint arXiv:2205.01068*, 2022.
- 625 Bo Zhao, Boya Wu, Muyang He, and Tiejun Huang. Svit: Scaling up visual instruction tuning.
626 *arXiv preprint arXiv:2307.04087*, 2023.
- 627 Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigt-4: En-
628 hancing vision-language understanding with advanced large language models. *arXiv preprint*
629 *arXiv:2304.10592*, 2023.
- 630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647

A ALGORITHM DETAILS

A.1 GREEDY CLUSTER MATCHING ALGORITHM

We present the pseudocode for the Greedy Cluster Matching in Algorithm 1.

Let $C_S = \{S_1, \dots, S_m\}$ and $C_C = \{C_1, \dots, C_n\}$ be cluster sets from SigLIP and CLIP embeddings respectively. The algorithm proceeds as follows:

1. Select the largest cluster from either set: $X_{max} = \arg \max_{X \in C_S \cup C_C} |X|$
2. If $X_{max} \in C_S$, find the best match in C_C : $Y_{best} = \arg \max_{C_j \in C_C} score(X_{max}, C_j)$
3. If $X_{max} \in C_C$, find the best match in C_S : $Y_{best} = \arg \max_{S_i \in C_S} score(X_{max}, S_i)$

Where the score function is defined as:

$$score(A, B) = \frac{|A \cap B|}{\frac{|A| + |B|}{2}}$$

This process is repeated, greedily selecting the largest remaining cluster and finding its best match, until all clusters are matched or one set is exhausted.

Algorithm 1 Greedy Cluster Matching Algorithm

Require: Two lists of clusters $c1$ and $c2$

Ensure: List of matched cluster pairs

- 1: $c1 \leftarrow \text{sort}(c1, \text{key} = \text{len}, \text{reverse} = \text{True})$
 - 2: $c2 \leftarrow \text{sort}(c2, \text{key} = \text{len}, \text{reverse} = \text{True})$
 - 3: $\text{matched_pairs} \leftarrow []$
 - 4: $\text{num_samples} \leftarrow 0$
 - 5: **while** $c1$ is not empty and $c2$ is not empty **do**
 - 6: **if** $\text{len}(c1[0]) \geq \text{len}(c2[0])$ **then**
 - 7: $\text{larger_cluster} \leftarrow c1.\text{pop}(0)$
 - 8: $\text{smaller_list} \leftarrow c2$
 - 9: **else**
 - 10: $\text{larger_cluster} \leftarrow c2.\text{pop}(0)$
 - 11: $\text{smaller_list} \leftarrow c1$
 - 12: **end if**
 - 13: $\text{best_match} \leftarrow \text{None}$
 - 14: $\text{best_score} \leftarrow -1$
 - 15: **for** $i, \text{cluster}$ in $\text{enumerate}(\text{smaller_list})$ **do**
 - 16: $\text{overlap} \leftarrow \text{len}(\text{set}(\text{larger_cluster}) \cap \text{set}(\text{cluster}))$
 - 17: $\text{avg_size} \leftarrow (\text{len}(\text{larger_cluster}) + \text{len}(\text{cluster}))/2$
 - 18: $\text{score} \leftarrow \text{overlap}/\text{avg_size}$
 - 19: **if** $\text{score} > \text{best_score}$ **then**
 - 20: $\text{best_score} \leftarrow \text{score}$
 - 21: $\text{best_match} \leftarrow (i, \text{cluster})$
 - 22: **end if**
 - 23: **end for**
 - 24: **if** best_match is not None **then**
 - 25: $\text{best_index}, \text{best_cluster} \leftarrow \text{best_match}$
 - 26: $\text{union} \leftarrow \text{list}(\text{set}(\text{larger_cluster}) \cup \text{set}(\text{best_cluster}))$
 - 27: $\text{matched_pairs}.\text{append}(\text{union})$
 - 28: $\text{num_samples} \leftarrow \text{num_samples} + \text{len}(\text{union})$
 - 29: $\text{smaller_list}.\text{remove}(\text{best_cluster})$
 - 30: **end if**
 - 31: **end while**
 - 32: **return** $\text{matched_pairs}, \text{num_samples}$
-

A.2 RANDOM SAMPLING WITH ITERATION

We present the pseudocode for the Random Sampling with Iteration in Algorithm 2.

Let $X = \{x_1, \dots, x_n\}$ be the set of embeddings.

k is a parameter that determines the power of the distance calculation (default to 12), and K is the desired number of selected embeddings.

1. Randomly select an initial embedding: $s_1 \in X$
2. Initialize selected set $S = \{s_1\}$
3. For $k = 2$ to K : $s_k = \arg \max_{x_j \in X \setminus S} \sum_{u \in S} \|x_j - x_u\|^k$ $S = S \cup \{s_k\}$
4. Return S

This formulation captures the process of iteratively selecting embeddings based on their cumulative distance from all previously selected embeddings, raised to the power k .

Algorithm 2 Random Sampling with Iteration

Require:

- 1: X : Set of embeddings
- 2: $num_samples$: Number of samples to select
- 3: k : Power factor for distance calculation (default: 12)

Ensure: Set of selected indices

- 4: $selected \leftarrow []$
 - 5: $n \leftarrow |X|$ ▷ Number of embeddings
 - 6: **for** $i = 1$ to $num_samples$ **do**
 - 7: $distances \leftarrow \text{zeros}(n)$
 - 8: **if** $selected$ is empty **then**
 - 9: $sampled_index \leftarrow \text{random_integer}(0, n - 1)$
 - 10: **else**
 - 11: **for** $j = 0$ to $n - 1$ **do**
 - 12: **if** $j \in selected$ **then**
 - 13: $distances[j] \leftarrow \infty$
 - 14: **else**
 - 15: $distances[j] \leftarrow \sum_{u \in selected} |||X[j] - X[u]|||^k$
 - 16: **end if**
 - 17: **end for**
 - 18: $inverted_distances \leftarrow \frac{1}{distances + \epsilon}$ ▷ ϵ is a small constant
 - 19: $distribution \leftarrow \frac{inverted_distances}{\sum inverted_distances}$
 - 20: $sampled_index \leftarrow \text{random_choice}(\text{range}(n), p = distribution)$
 - 21: **end if**
 - 22: $selected.append(sampled_index)$
 - 23: **end for**
 - 24: **return** $selected$
-


B DATA SAMPLES

For the sake of brevity, we have included only two of the numerous multi-turn interactions present in each data sample.

B.1 GREEDY CLUSTER MATCHING ALGORITHM

Samples obtained through Greedy Cluster Matching typically feature similar subjects and shot compositions, but when paired with carefully crafted prompts, these similarities can be leveraged to generate more challenging and nuanced questions.

756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809




...clear day, two white sheep are peacefully grazing on a verdant hillside is facing away from us...

...sheep stands in the focal point of the image...the sheep with its body painted...


...the sheep, with its wool as white as a cloud, is facing the camera, its gaze meeting ours...

...sheep on the left is facing away from the camera, its attention seemingly focused on the grass...

 How do the images collectively convey a sense of rural life, and what can be inferred about the relationship between nature and the animals that inhabit it?

The images showcased a sense of rural life by showcasing serene and beautiful scenes of animals in their natural habitats...The relationship between nature and the animals...

...

 How many animals are in all the images?

To determine the total of animals in all the images, we need to count the animals in each image. Image 1 features 2 sheep, Image 2 features 1 sheep...Answer = 6.



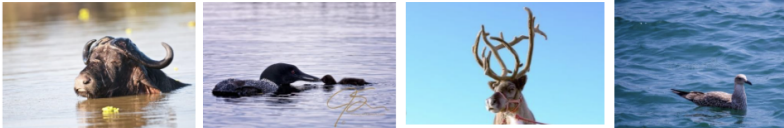



Figure 6: Images sampled from the same matched cluster often feature similar subjects or scenes.

B.2 RANDOM SAMPLING WITH ITERATION

Random sampling tends to yield greater diversity in subjects compared to Greedy Clustering, striking a balance between variety and relatedness that can potentially lead to more robust and wide-ranging question sets.



...a black water buffalo is captured...with large, curved horns is partially submerged...

...a mother loon and her two chicks...a majestic creature with a black body...

...the reindeer, with its coat of brown and white, is adorned with a vibrant red harness...

...the bird, with its wings neatly folded at its sides, floats gently on surface...



 'Which image best represents a moment of serenity and solitude, and why?

Image 1 best represents a moment of serenity and solitude. The image features a black buffalo partially submerged in calm water, with its eyes gently closed ...The presence of yellow flowers in the distance adds a touch of color to the scene...

...

 What is the common thread among the subjects in the images, and how do their environments contribute to the overall mood?

The common thread among the subjects in the images is that they are all animals in their natural habitats. The environments in each image contribute to the overall mood by creating a sense of tranquility and peacefulness....






Figure 7: Images sampled using the iterative algorithm allow for different yet related subjects (e.g., various animal species)

810 C PROMPT

811
812 While prompts play a crucial role in data generation, optimizing them remains a significant chal-
813 lenge. After numerous iterations, we identified two particularly effective prompts for multi-image
814 data generation.

816 C.1 LLaVA STYLE PROMPT

817
818 Inspired by LLaVA (Liu et al., 2024b), our approach utilizes a specialized prompt to address simpler
819 multi-image and single-image tasks, focusing on more straightforward visual comprehension and
820 analysis.

821
822
823 *You are an AI visual assistant capable of analyzing multiple images, including*
824 *both visual content and textual elements using Optical Character Recognition*
825 *(OCR). You will receive four to five images, each potentially accompanied by*
826 *captions and containing text, numbers, signs, or other recognizable characters.*
827 *Your task is to create a plausible and challenging question that involves*
828 *comparison, ranking, storytelling, logical reasoning, or detailed textual analysis*
829 *across the images, and then provide a detailed answer...*

830
831 Figure 8: LLaVA-style prompt for OCR and smaller visual task data generation

833 C.2 LONGER PROMPT

834
835 Our approach aims to generate more complex, multi-turn questions that require in-depth reasoning
836 across multiple images.

837
838
839 *Create questions that ask to compare elements across the images, such as*
840 *identifying which image best represents a critical or turning point moment, quality,*
841 *or characteristic; formulate questions that require ranking the images based on*
842 *intricate and plausible criteria, such as strategic importance, sequence, or visual*
843 *impact; develop questions that involve piecing together a narrative from the images,*
844 *understanding a sophisticated sequence of events, or explaining a complex*
845 *progression shown; and ask questions that require advanced logical reasoning to*
846 *deduce why certain elements are present, the purpose behind actions shown, or the*
847 *broader implications of what is depicted...*

848
849 Figure 9: SMIR prompt to generate more complex question and answers

850
851
852
853
854
855
856
857
858
859
860
861
862
863