Vision-Flan: Scaling Human-Labeled Tasks in Visual Instruction Tuning

Zhiyang Xu[♠] **Chao Feng**[♣] **Rulin Shao**[♡] **Trevor Ashby**[♠] **Ying Shen**[♠]
Di Jin[◇] **Yu Cheng**[♦] **Qifan Wang**[◇] **Lifu Huang**[♠]
[♠]Virginia Tech [♡]University of Washington [♣]University of Michigan
[♦]The Chinese University of Hong Kong [◇]Meta AI
[♠]{zhiyangx,trevorashby,yings,lifuh}@vt.edu
[♣]chfeng@umich.edu [♡]rulins@cs.washington.edu [◇]jindi@meta.com
[♦]chengyu@cse.cuhk.edu.hk [◇]wqfcr@meta.com

Abstract

Despite vision-language models' (VLMs) remarkable capabilities as versatile visual assistants, two substantial challenges persist within the existing VLM frameworks: (1) *lacking task diversity* in pretraining and visual instruction tuning, and (2) *annotation error* and *bias* in GPT-4 synthesized instruction tuning data. Both challenges lead to issues such as poor generalizability, hallucination, and catastrophic forgetting. To address these challenges, we construct VISION-FLAN, the most diverse publicly available visual instruction tuning dataset to date, comprising 187 diverse tasks and 1,664,261 instances sourced from academic datasets, and each task is accompanied by an expert-written instruction. In addition, we propose a two-stage instruction tuning framework, in which VLMs are firstly finetuned on VISION-FLAN and further tuned on GPT-4 synthesized data. We find this two-stage tuning framework significantly outperforms the traditional single-stage visual instruction tuning framework and achieves the state-of-the-art performance across a wide range of multi-modal evaluation benchmarks. Finally, we conduct in-depth analyses to understand visual instruction tuning and our findings reveal that: (1) GPT-4 synthesized data does not substantially enhance VLMs' capabilities but rather modulates the model's responses to human-preferred formats; (2) A minimal quantity (e.g., 1,000) of GPT-4 synthesized data can effectively align VLM responses with human-preference; (3) Visual instruction tuning mainly helps large-language models (LLMs) to understand visual features. Our dataset and models are publicly available at `https://github.com/VT-NLP/Vision-Flan`.

1 Introduction

Recent vision-language models (VLMs) (Liu et al., 2023e; Li et al., 2023d; Dai et al., 2023), built upon pre-trained large-language models (LLMs) (Chiang et al., 2023; Gao et al., 2023) and pretrained image encoders (Sun et al., 2023), have shown impressive capabilities as general visual assistants. Besides the unimodal encoders, the main ingredients of these VLM frameworks encompass: (1) a bridging module, such as the MLP layers in the LLaVA model (Liu et al., 2023e; Li et al., 2023d), that establishes connections between the pretrained image encoders and LLMs, (2) large-scale text-image pairs (Schuhmann et al., 2022) used for pre-training the bridging module, and (3) GPT-4 synthesized visual instruction tuning datasets (Liu et al., 2023e; Li et al., 2023b) to align the responses of VLMs with human preferences (i.e., following users' instruction to generate detailed and helpful responses). Despite their notable successes, we identify two remaining challenges that merit further investigation.

Firstly, the data used in the pre-training stage is dominated by the image captioning task, which lacks diversity, resulting in limited generalizability of VLMs (Chen et al., 2023c; Zhang et al., 2023). For instance, the LLaVA model (Liu et al., 2023e) performs poorly on the optical character recognition (OCR) task due to the absence of instances related to text detection during pre-training (Zhang et al., 2023). Several recent studies address this problem by further fine-tuning VLMs on instruction tuning datasets covering more tasks (Zhang et al., 2023; Hu et al., 2023; Liu et al., 2023d) such as visual question answering and OCR but the coverage of the tasks is still limited.

Secondly, most of the existing visual instruction tuning datasets (Liu et al., 2023e; Li et al., 2023b; Yin et al., 2023) are synthetically generated via GPT-4 by repurposing text annotations such as captions or dense captions from existing computer-vision datasets to generate new tasks, such as visual dialogue, Complex VQA and detail captions. While they enable VLMs to generate fluent and detailed responses aligned with human preferences, the lack of task diversity, spurious co-occurring pat-

Findings of the Association for Computational Linguistics: ACL 2024, pages 15271–15342
August 11-16, 2024 ©2024 Association for Computational Linguistics

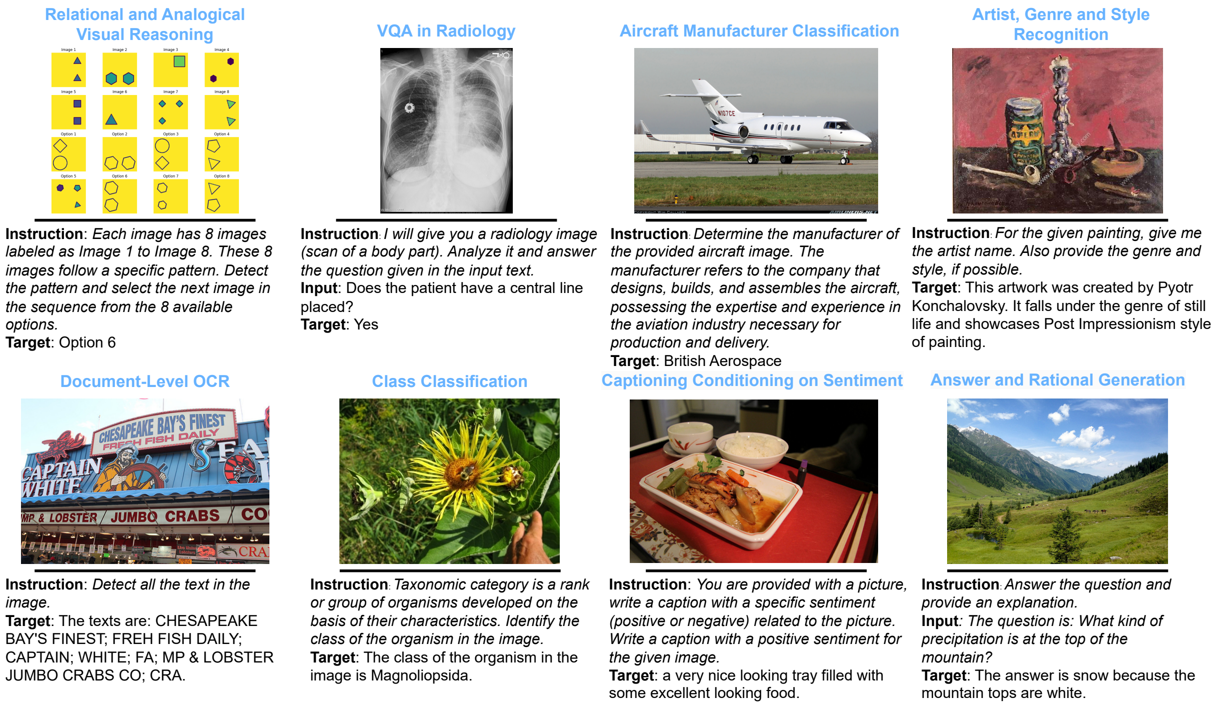


Figure 1: Sample tasks in VISION-FLAN. **Instruction** denotes a task instruction crafted by annotators. **Input** means text input in the given task, and **Target** is the target response based on the instruction.

terns between objects, and long-form outputs may cause severe hallucination (Liu et al., 2023c; Li et al., 2023g; Liu et al., 2023a; Zhou et al., 2023), and catastrophic forgetting – VLMs fail to maintain a similar classification performance on basic detection tasks, such as MNIST (LeCun, 1998) and CIFAR-10 (Krizhevsky et al., 2009), compared to the zero-shot performance of their vision encoders (Zhai et al., 2023).

To address both challenges, we introduce VISION-FLAN, the most diverse public-available visual instruction tuning dataset consisting of 187 tasks drawn from academic datasets, covering *perception* tasks such as object detection and OCR, *domain-specific* tasks such as image-quality classification and image-style classification, *complex reasoning* tasks such as graph understanding and geometric question answering, and many more. Each task in VISION-FLAN is accompanied by an expert-written instruction. We show some sample tasks from VISION-FLAN in Figure 1 and provide the full list of tasks in Appendix J.

In addition, we introduce a two-stage instruction tuning framework. In the first stage, we utilize the pre-trained LLaVA model (Liu et al., 2023e) as our initial model, and finetune it on VISION-FLAN to gain diverse capabilities, resulting in the VISION-FLAN BASE model. However, due to the concise nature of target outputs in academic datasets, the responses generated by VISION-FLAN BASE tend to be brief and not aligned with human preferences. Therefore, in the second stage, we further finetune VISION-FLAN BASE using a minimal amount of GPT-4 synthesized data. This step aims to adjust the model's outputs to be more in line with human preferences, resulting in the VISION-FLAN CHAT model.

Our experimental results demonstrate that high-quality human annotations from VISION-FLAN significantly enhance the capabilities of both VISION-FLAN BASE and VISION-FLAN CHAT while reducing the risk of hallucination and catastrophic forgetting. The two-stage instruction tuning framework enables VISION-FLAN CHAT to achieve better human preference alignment using much less GPT-4 synthesized data compared to the state-of-the-art VLMs. Finally, we perform extensive analysis to understand visual instruction tuning including the roles of human-labeled and GPT-4 synthesized data, and the impacts of various training strategies. Our investigation yields several key insights:

- Increasing the number of human-labeled tasks in visual instruction tuning can substantially enhance VLMs' capabilities across extensive evaluation benchmarks.
- GPT-4 synthesized data does not substantially enhance VLMs capabilities and yields marginal improvements in the VLMs' perfor-

mance on comprehensive evaluation benchmarks, such as MME (Fu et al., 2023) and MM-Bench (Liu et al., 2023f).
- A minimal quantity (1,000) of GPT-4 synthesized data is sufficient to align VLMs' responses with human preferences. Notably, increasing GPT-4 synthesized data does not correspond to a proportional enhancement in alignment and introduces hallucination and bias into the VLMs.
- Visual instruction tuning mainly enhances the ability of large-language models (LLMs) to process and understand visual features. The training of the bridging module, which maps visual features to the embedding space of LLMs, is predominantly achieved during the pre-training phase.

2 VISION-FLAN

2.1 Collection Pipeline

We carefully design an annotator selection process to identify qualified annotators, which involves 2 iterations of training and testing. More details of the selection process and compensation can be found in Appendix A.1. In the end, we hire 7 out of 21 candidates as our annotators and all of them are graduate students in computer science. To ensure the diversity and quality of the tasks in VISION-FLAN, we design a rigorous annotation pipeline with four major steps:

Existing dataset collection and pre-processing: Two expert researchers (i.e., senior Ph.D. students in the fields of natural language processing and computer vision) search online and identify high-quality vision-language datasets. The datasets are then equally distributed to 7 annotators to download and preprocess the datasets. Each processed instance consists of an image, an instruction (the task definition from the original dataset with minor modifications), a text input if applicable, and a target output.

Creating new tasks: The two expert researchers and annotators also discuss potential new tasks that could be derived from the existing annotations. We derive new tasks by combining the annotations of two or more existing tasks on a dataset. For example, in the Concadia dataset (Kreiss et al., 2022), each instance consists of an image caption and a knowledge snippet related to the image. We propose a new task to predict both the caption and the background knowledge given an image, which is a free-form generation task. The new target output is formed by concatenating the caption with the knowledge snippet. We also develop new tasks by creating more basic versions of the original tasks. For example, given the object detection annotations in MSCOCO (Lin et al., 2014), we propose an object selection task in which we provide a list of objects and ask the model to select the object that appears in the image (the negative options are created by sampling objects that appear in other images but not in the given image). The expert researchers and annotators manually solve 20 instances for each newly developed task. If the human predictions match the target outputs, this new task is considered valid.

Iteratively refining the task instructions and output templates: For existing tasks, we ask annotators to write instructions based on the original task definitions with minor modifications. For newly developed tasks, the annotators write instructions by discussing with the expert researchers. Once an annotator finishes writing a new instruction, one of the two expert researchers is randomly assigned to examine the instances and provide feedback for revising the instruction. This step iterates repeatedly until the instruction meets our requirements. We require the instruction to be *clear*, *easy to understand*, and can *be correctly executed by a human*. Each task together with its associated dataset and instruction is then added to the pool of candidate tasks for VISION-FLAN.

Verifying the quality of each task: From the candidate task pool, two expert researchers, including a native English speaker, work together to select the high-quality tasks where the instruction is fluent and effectively conveys the intended task and the task does not overlap with other tasks.

Based on these four steps, we finally collect 187 high-quality tasks, and for each task, we randomly sample 10,000 instances from its corresponding dataset. If a dataset contains less than 10,000 instances, we include all of them. We name the dataset as VISION-FLAN, consisting of 1,664,261 instances for 187 tasks in total. We include references to all the datasets used in VISION-FLAN in Appendix H and show an instance for each task in Appendix J.

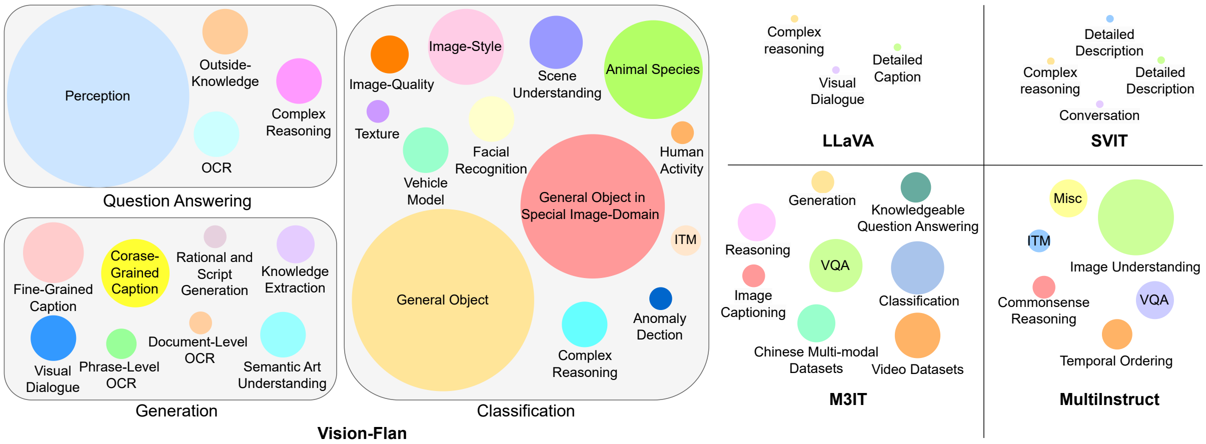


Figure 2: Comparison of task diversity between VISION-FLAN and previous visual instruction tuning datasets. LLaVA and SVIT report very coarse-grained categories of tasks. Each circle represents a task category and the radius is proportional to the number of tasks in that category. The radius of circles for different datasets are comparable.

Dataset	Instances #	Tasks #	Source
LLaVA (Liu et al., 2023e)	150K	3	Synthetic
LAMM (Yin et al., 2023)	196K	8	Synthetic
VL-Qwen (Bai et al., 2023a)	350K	Unknown	Private
M^3IT (Li et al., 2023e)	2.4M	40	Synthetic
mPlug-Owl (Ye et al., 2023)	150K	3	Synthetic
Shikra (Chen et al., 2023a)	156K	4	Synthetic
SVIT (Zhao et al., 2023)	4.2M	4	Synthetic
MultiInstruct (Xu et al., 2023)	510K	62	Public
VISION-FLAN (Ours)	1.6M	196	Public

Table 1: Comparison between VISION-FLAN and existing visual instruction tuning datasets.

2.2 Comparison with Existing Datasets

Table 1 presents a comparison between existing visual instruction tuning datasets and VISION-FLAN. For existing visual instruction tuning datasets, we directly adopt the numbers of tasks and instances reported in their original papers. The majority of these datasets are generated using proprietary language models, such as ChatGPT[1] and GPT-4[2], and exhibit a narrow range of task diversity. VL-Qwen (Bai et al., 2023a) is a recently introduced large-scale dataset annotated by humans but remains inaccessible to the public. Although MultiInstruct (Xu et al., 2023) is based on publicly available datasets, it mainly focuses on visual grounding tasks and only contains 29 tasks that do not involve region-specific information. In contrast, VISION-FLAN encompasses a significantly more diverse array of tasks, offering a three-times increase compared to the number of tasks in MultiInstruct.

In Figure 2, we compare the task categories covered by VISION-FLAN and other datasets. Tasks within VISION-FLAN are first categorized into three primary groups: *Question Answering*, *Classification*, and *Generation*, and each of these primary groups is further divided into specific, fine-grained categories. For instance, within the *Classification* group, the *General Object* category involves classifying objects in images into various concepts, such as “fish”, “car”, and “dog”. Contrastingly, the *Vehicle Model* category demands the models to accurately identify specific car brands or models, like “Toyota” and “Camry”. The visualization in Figure 2 clearly demonstrates the superior diversity and volume of tasks in VISION-FLAN compared to existing datasets. We list tasks in each category in Appendix I.

3 VISION-FLAN Finetuning

Model Architecture We adopt the same VLM architecture as LLaVA (Liu et al., 2023d) and denote it as LLaVA-Architecture. As shown in Figure 3, it consists of a pre-trained vision encoder, a pre-trained large language model, and two layers of MLPs to connect them. In the vision-language pre-training phase of the LLaVA-Architecture, both the pre-trained vision encoder and large language model remain frozen, and only the MLP layers are trained on a large-scale image captioning dataset (Schuhmann et al., 2022). We leverage this pre-trained LLaVA model , without any visual instruction tuning, as our initial model and finetune it on VISION-FLAN. During visual instruction tuning, we finetune both the MLP layers and the language model while keeping the vision encoder frozen.

Two-stage Visual Instruction Tuning Contrary to prior approaches (Liu et al., 2023d; Dai et al.,

[1] https://openai.com/blog/chatgpt
[2] https://openai.com/research/gpt-4

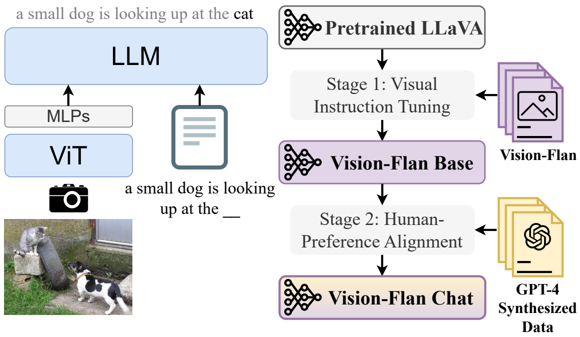


Figure 3: The left of the figure shows the LLaVA-Architecture and the right of the figure shows the two-stage visual instruction tuning pipeline.

2023) that mix human-labeled data with GPT-4 synthesized data for visual instruction tuning, our study introduces a two-stage instruction tuning pipeline. As shown in Figure 3, in the first stage, we finetune the VLM on all 187 tasks of VISION-FLAN to acquire diverse capabilities and name the resulting model as VISION-FLAN BASE. However, due to the brevity of target outputs presented in academic datasets, the responses from VISION-FLAN BASE are not in human-preferred formats. Hence, we further finetune VISION-FLAN BASE on GPT-4 synthesized data to align the model's outputs with human preference. We denote the yielded model as VISION-FLAN CHAT. This training framework requires minimal GPT-4 synthesized data while providing deep insights into the distinct contributions of human-labeled and GPT-4 synthesized data in visual instruction tuning.

Implementation Details We leverage LLaVA-Architecture with Vicuna-13B v1.5 (Chiang et al., 2023), CLIP-ViT-L-336px (Radford et al., 2021) and two layers of MLP as our VLM. For the first-stage instruction tuning, we finetune the MLP layers and the language model on VISION-FLAN for 1 epoch with a learning rate 2e-5 and per device batch size 16 on 8 A100 GPUs. For the second-stage instruction tuning, we further finetune the MLP layers and the language model on 1,000 instances randomly sampled from the LLaVA dataset (Liu et al., 2023e) with learning rate 1e-5 and per device batch size 8 on 8 GPUs for 128 steps. In the following sections, we use LLaVA dataset and GPT-4 synthesized data interchangeably.

4 Experiment Setup

Evaluation Datasets We evaluate the models on several widely adopted multimodal evaluation benchmark datasets including *multiple-choice* benchmarks: **MMbench** (Liu et al., 2023f), **MME** (Fu et al., 2023), and **MMMU**; *free-form generation* benchmarks: **MM-Vet** (Yu et al., 2023) and **LLaVA-Bench**; the *hallucination* benchmark: **POPE** (Li et al., 2023g), and *catastrophic forgetting* benchmarks: **CIFAR-10 and CIFAR-100** (Krizhevsky et al., 2009), **MNIST** (LeCun, 1998), and **miniImageNet** (Vinyals et al., 2016). More details of the evaluation datasets can be found in Appendix B.

Evaluation Protocols For MMbench, MME, MM-Vet, LLaVA-Bench, POPE and MMMU, we strictly follow their official implementations of evaluation code to evaluate the performance of each model. For datasets that do not have official evaluation codes including CIFAR-10, CIFAR-100, MNIST, and miniImageNet, we leverage the state-of-the-art open-source LLM, Vicuna 1.5 13B, to perform the evaluation and report the averaged performance on these four datasets in the CF column in Table 2. More details of evaluation protocols can be found in Appendix C.

Baselines We compare our models with several recent state-of-the-art vision-language models, including **BLIP-2** (Li et al., 2023d), **InstructBLIP** (Dai et al., 2023), **Shikra** (Chen et al., 2023a), **LLaVA** (Liu et al., 2023e), **Qwen-VL**, **Qwen-VL-Chat** (Bai et al., 2023b), and **LLaVA-1.5** (Liu et al., 2023d). The LLMs and image encoders used in all baselines are shown in Table 2. Details of baselines can be found in Appendix D.

5 Results and Analysis

5.1 Main Results

As demonstrated in Table 2, VISION-FLAN BASE achieves state-of-the-art performance on comprehensive evaluation benchmarks including MME, MM-Bench and MMMU, while reducing hallucination and catastrophic forgetting. However, we observe that VISION-FLAN BASE scores significantly lower on the LLaVA-Bench dataset in comparison to VLMs trained using GPT-4 synthesized data. We attribute this discrepancy to the conciseness and brevity of target outputs within academic datasets. As shown in Figure 1, VQA tasks frequently yield outputs comprising a single or a few words. Even outputs of many generation tasks are typically confined to one or two succinct sentences. Training on these tasks leads VISION-FLAN BASE to generate brief responses, which are not aligned with human

Model	LLM	Image Encoder	MM-Bench	MME	MMMU	LLaVA-Bench	MM-Vet	Pope	CF
BLIP-2	FlanT5-XXL	ViT-g/14	-	1293.8	34.0	-	22.4	85.3	-
InstructBlip	Vicuna-13B	ViT-g/14	36.0	1212.8	33.8	58.2	25.6	78.9	-
Mini-GPT4	Vicuna-13B	ViT-g/14	24.3	581.67	27.6	-	-	-	-
Shikra	Vicuna-13B	ViT-L/14	58.8	-	-	-	-	-	-
LLaVA	Vicuna-13B v1.5	CLIP-ViT-L-336px	38.7	1151.6	-	70.8	33.4	75.3	-
Qwen-VL	Qwen-7B	ViT-bigG	38.2	-	-	-	-	-	-
Qwen-VL-Chat	Qwen-7B	ViT-bigG	60.6	1487.5	32.9	73.6	-	-	72.1
LLaVA 1.5	Vicuna-13B v1.5	CLIP-ViT-L-336px	66.7	1531.3	33.6	70.7	35.4	83.6	73.3
VISION-FLAN BASE	Vicuna-13B v1.5	CLIP-ViT-L-336px	**69.8**	**1537.8**	**34.4**	38.5	33.4	85.9	**87.2**
Second-Stage Tuning with 1,000 GPT-4 Synthesized Instances									
VISION-FLAN CHAT	Vicuna-13B v1.5	CLIP-ViT-L-336px	67.6	1490.6	34.3	**78.3**	**38.0**	**86.1**	84.0

Table 2: Comprehensive evaluation of VLMs on widely adopted benchmark datasets. CF denotes the averaged performance of VLMs on four catastrophic forgetting benchmarks.

preferences. Conversely, through the second-stage tuning on a mere 1,000 GPT-4 synthesized data instances, VISION-FLAN CHAT achieves significant performance improvement on LLaVA-Bench, a benchmark measuring human-preference alignment, while maintaining a relatively lower rate of hallucination and catastrophic forgetting. To better understand why VISION-FLAN models are better than current VLMs, we conduct two case studies focusing on OCR and Entity Recognition and analyze both quantitative and qualitative results in Appendix E.2.

Another finding in Table 2 is that compared to VISION-FLAN BASE, VISION-FLAN CHAT achieves slightly inferior performance on comprehensive evaluation benchmarks demonstrating the bias and hallucination inevitably introduced by the GPT-4 synthesized data, which is discussed in detail in Section 5.2.

5.2 Effect of Human-Labeled and GPT-4 Synthesized Datasets

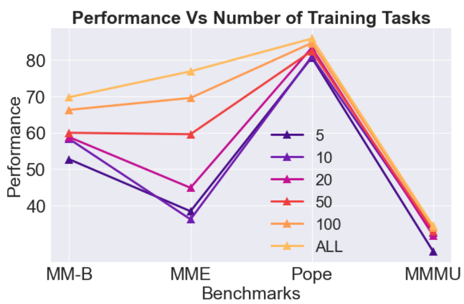


Figure 4: Performance on four comprehensive benchmarks versus the number of training tasks.

Effect of Task Diversity in VISION-FLAN Figure 4 illustrates the relationship between the number of tasks from VISION-FLAN employed during visual instruction tuning and the performance of VISION-FLAN BASE across four comprehensive evaluation benchmarks. It's apparent that as the number of tasks increases, the performance of VISION-FLAN BASE on all datasets is improved. To evaluate the impact of varying numbers of instances from different tasks, we fix the total amount of instances used for visual instruction tuning and experiment with different numbers of tasks. As demonstrated in Table 3, when the number of training instances is constant, augmenting the number of tasks significantly enhances model performance. These findings substantiate our hypothesis that *the diverse array of human-labeled tasks within* VISION-FLAN *is essential for improving the capabilities of VLMs*.

# of Tasks	# of Instances per Task	MMB	MME	Pope	MMMU
Training with 100,000 Instances					
10	10,000	58.3	723.9	81.0	32.6
187	500	58.8	1314.3	83.3	33.3
Training with 200,000 Instances					
20	10,000	58.8	897.3	83.4	31.8
187	1,000	63.5	1373.5	83.6	33.7

Table 3: Comparison of VISION-FLAN BASE trained with a fixed total amount of data instances.

Effect of GPT-4 Synthesized Data on Comprehensive Evaluation Benchmarks Furthermore, we analyze if GPT-4 synthesized data can improve the model's performance on comprehensive evaluation benchmarks and show the results in Figure 5. Further tuning VISION-FLAN BASE on GPT-4 synthesized data instances does not lead to performance improvement. Tuning pretrained LLaVA model on a small amount of GPT-4 synthesized data (100) can improve its performance on MME but further increasing the number of training instances does not lead to any improvement. We also observe a similar trend on the MM-Bench dataset and report the result in Appendix E.1. These observations are in line with recent findings in LLMs: *GPT-4 synthesized data does not improve model's* *capability but rather modulates the responses towards human-preferred formats* (Jain et al., 2023; Gudibande et al., 2023).

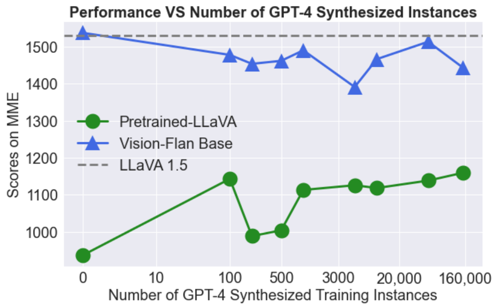


Figure 5: Effect of the number of GPT-4 synthesized training instances on MME. The dashed gray line indicates the performance of LLaVA 1.5.

Effect of GPT-4 Synthesized Data on Human-Preference Alignment When utilizing our proposed two-stage tuning framework, we find that by performing a second-stage finetuning on a mere 1,000 GPT-4 synthesized instances from the LLaVA dataset, VISION-FLAN CHAT achieves significantly better performance (78.5 v.s. 38.5) on the LLaVA-Bench dataset. This observation leads us to raise the question: *Is extensive finetuning on large-scale GPT-4 synthesized datasets necessary for aligning VLMs with human preferences?* To answer it, we finetune both VISION-FLAN BASE and pretrained LLaVA model on different numbers of GPT-4 synthesized instances ranging from 100 to 158,000, and show the results in Figure 6. As we can see, with 1,000 instances, VISION-FLAN BASE achieves a score of 78.3 and further increasing the number of training instances leads to a performance drop. A similar trend can also be seen for the pretrained LLaVA model.

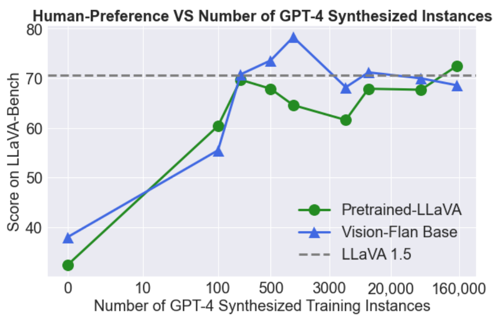


Figure 6: Effect of the number of GPT-4 synthesized instances on human preference alignment. The dashed gray line indicates the performance of LLaVA 1.5.

GPT-4 Synthesized Data Causes Hallucination and Bias Concurrent work (Liu et al., 2023c) identifies that hallucination in current VLMs can be caused by their bias toward positive answers (i.e., "Yes"). In Figure 7, we explicitly show the relationship between hallucination, the ratio of "Yes", and the number of training instances from GPT-4 synthesized dataset. As the number of GPT-4 synthesized instances increases, the model's responses are biased towards the answer "Yes" even if the objects are not in the images, causing the model to hallucinate. This observation suggests that a small amount of GPT-4 synthesized training instances is preferred to avoid hallucination and bias in VLMs.

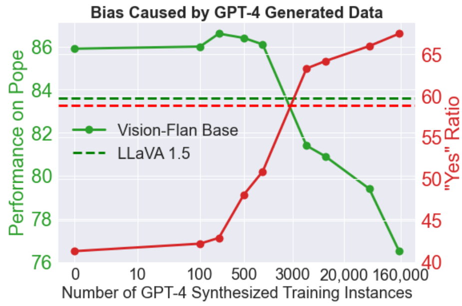


Figure 7: Effect of the number of GPT-4 synthesized training instances on the hallucination benchmark and the ratio of "Yes". The dashed lines indicate the performance of the state-of-the-art LLaVA 1.5 model.

5.3 Single-stage Tuning on Mixed Data Vs. Two-stage Tuning

In this section, we compare the performance of two training strategies based on the same pretrained LLaVA model: (1) finetuning it on the mix of VISION-FLAN and the LLaVA dataset; (2) finetuning it utilizing VISION-FLAN and 1,000 instances from the LLaVA dataset with our two-stage tuning method. As illustrated in Table 4, the performance of VLMs finetuned on the mix of VISION-FLAN and GPT-4 synthesized data is notably inferior compared to VISION-FLAN CHAT trained through our two-stage tuning framework.

Method	# of LLaVA	MME	LLaVA-Bench	MM-Vet
Mixed Data	1,000	1364.0	52.7	36.6
Mixed Data	158,000	1317.9	63.9	36.8
Two-stage	1,000	1490.6	78.3	38.0

Table 4: Comparison between single-stage finetuning on mixed data and two-stage finetuning.

5.4 Effect of Newly Created Tasks

In the data collection phase of VISION-FLAN, we collaborate with annotators to derive 65 novel tasks from pre-existing annotations. To evaluate the impact of these newly introduced tasks, we conducted an experiment where a VISION-FLAN BASE was trained exclusively on the existing tasks and its performance was compared against the VISION-FLAN BASE trained on all tasks, including the new additions. The comparative results are presented in Table 5. The outcomes distinctly demonstrate the advantages of expanding the task set through the utilization of existing annotations.

Model Name	MME	MM-Bench	Pope
VISION-FLAN BASE	1537.8	69.8	85.9
VISION-FLAN BASE w/o New Tasks	1379.8	67.3	84.6

Table 5: Comparison between finetuning VISION-FLAN BASE on all tasks and finetuning VISION-FLAN BASE only on existing tasks.

5.5 Contributions of Tasks from Different Task Groups

Understanding the contribution of each task to the models' performance is crucial for visual instruction tuning. However, training a model on each task in VISION-FLAN can impose significant computational cost. Instead, we propose a group-level analysis of task contributions. As illustrated in Figure 2, tasks are categorized into three primary groups: Question Answering (QA), Classification, and Generation. We train three VISION-FLAN CHAT models on VISION-FLAN, each time excluding all tasks from one primary group, and show their performance in Table 6.

Model Name	MME	Pope	LLaVA-Bench	MM-Vet
VISION-FLAN CHAT	1537.8	86.1	78.3	38.0
w/o QA	1211.6	83.3	73.5	37.9
w/o Classification	1376.6	84.9	75.4	37.3
w/o Generation	1390.9	83.8	68.1	36.2

Table 6: Contributions of different tasks group to the performance of VISION-FLAN CHAT.

Our findings indicate that: (1) the inclusion of generation tasks markedly enhances the model's capability in free-form generation tasks, as evidenced by performance on LLaVA-Bench; (2) QA tasks and classification tasks contribute significantly to model's performance MME, and (3) synergistic interactions between different task groups lead to further performance enhancements.

5.6 What is Essentially Improved in VLMs during Visual Instruction Tuning

In LLaVA-Architecture, the MLP layers map the visual features from a vision encoder into the embedding space of LLMs. The LLMs then interpret the visual features and follow text instructions to generate responses. In Table 7, we show the results of training different modules during visual instruction tuning and observe that solely tuning MLPs causes a significant performance drop compared to tuning both MLPs and LLMs during visual instruction tuning. However, tuning LLMs with frozen MLPs results in similar performance as tuning both modules, demonstrating that visual instruction tuning mainly enables LLMs to better understand visual features while MLPs have been sufficiently learned during pretraning. To further support this claim, we replace the instruction-tuned MLPs in VISION-FLAN BASE and VISION-FLAN CHAT with the pretrained MLPs from the pre-trained LLaVA model, and show that with the pretrained MLPs, both models can retain more than 90% of performance on most tasks as shown in Table 8. We also compute the Pearson Correlation Coefficient between the parameters of pretrained MLPs and instruction-tuned MLPs, and find that their correlation coefficient is higher than 0.99.

LLM	MLPs	MM-Bench	MME	LLaVA-Bench	Pope
✗	✗	45.0	936.3	32.4	51.9
✗	✓	52.4	1107.3	39.1	83.3
✓	✗	69.2	1495.5	39.3	85.6
✓	✓	69.8	1537.8	38.5	85.9

Table 7: Effect of tuning different modules in VISION-FLAN BASE. ✓ denotes the module is tuned and ✗ denotes the module is frozen during visual instruction tuning.

Model	MMB	MME	LLaVA-Bench	Pope
VISION-FLAN BASE	69.8	1537.8	38.5	85.9
+ Pretrained MLP	68.0	1403.1	36.4	84.0
VISION-FLAN CHAT	67.6	1490.6	78.3	86.1
+ Pretrained MLP	65.7	1332.2	73.8	85.4

Table 8: Results of replacing visual instruction tuned MLPs with pretrained MLPs. Gray rows show the performance of the original models and yellow rows show the performance after replacing instruction-tuned MLPs with pretrained MLPs.

6 Related Work

Instruction tuning (Wei et al., 2022) is first introduced in NLP and has been adapted to the visual-language domain. MultiInstruct (Xu et al., 2023) propose the first human-label multi-modal instruction tuning dataset for improving the zero-shot performance of pre-trained VLMs. LLaVA (Liu et al., 2023e) leverage GPT-4 to repurpose text annotations such as captions or dense captions from existing computer-vision datasets to generate visual dialogues, Complex VQA and detail captions for visual instruction tuning. Following LLaVA, mPLUG-Owl (Ye et al., 2023), LAMM (Yin et al., 2023), MIMIC-IT (Li et al., 2023a) and Macaw-LLM (Lyu et al., 2023) leverage proprietary LLMs such as GPT-4 and ChatGPT to further extend the instruction tuning tasks into 3D-domain, multiple-images and videos, and increase the amount of training instances. MiniGPT-4 (Zhu et al., 2023) utilizes ChatGPT to refine output from the pre-trained VLM itself. InstructBLIP (Dai et al., 2023) and LLaVA-1.5 (Liu et al., 2023d) mix the human-annotated and GPT4 synthesized datasets to enhance visual instruction tuning.

Several recent work explores different strategies to improve visual instruction tuning. StableLLaVA (Li et al., 2023f) and VPG-C (Li et al., 2023c) generate both images and texts using Stable Diffusion (Rombach et al., 2022) or Blended Diffusion (Avrahami et al., 2022) to alleviate domain bias and encourage VLMs attend to visual details. (Liu et al., 2023b) demonstrate the bias introduced by positive instructions and introduce negative instruction examples for improving robustness. Shikra (Chen et al., 2023a) incorporate visual grounding tasks in visual instruction tuning to improve the VLM's referential capability. LLaVAR (Zhang et al., 2023) and BLIVA (Hu et al., 2023) leverage OCR tools and GPT-4 to generate tasks helping VLMs to understand text in images. (Lu et al., 2023) and SVIT (Zhao et al., 2023) empirically study the effect of scaling the size of VLMs and the size of GPT-4 synthesized dataset. Two concurrent works (Wang et al., 2023a; Chen et al., 2023b) directly prompt GPT-4V with images as input to generate visual instruction tuning data and achieve superior performance. Additional related work can be found in Appendix G.

Unlike all prior work, our work mainly focuses on scaling human-labeled tasks in visual instruction tuning to improve VLMs' capabilities. Additionally, we perform extensive analysis to understand the characteristics of human-labeled and GPT-4 synthesized data and draw meaningful conclusions.

7 Conclusion

We construct VISION-FLAN, the most diverse public-available visual instruction tuning dataset, consisting of 187 diverse tasks and 1,664,261 instances collected from academic datasets, and each task is accompanied by an expert-written instruction. We demonstrate that VLMs trained on VISION-FLAN with proposed two-stage tuning framework achieve state-of-the-art performance on comprehensive evaluation benchmarks. Additionally, we perform extensive analysis and reveal the distinct contributions of human-labeled and GPT-4 synthesized data in visual instruction tuning.

8 Limitations

All the tasks included in VISION-FLAN are in English, which confines the usage of our dataset and models to English speaking populations. Future work should extend VISION-FLAN with multi-lingual tasks. In addition, all the tasks in VISION-FLAN only consists of a single image. Many real-world vision-language tasks require the model to take multiple images as inputs, such as script generation (Qi et al., 2024) and multimodal entity linking (Yao et al., 2024). Thus, future work should explore vision-language tasks that involve multiple images or videos.

Our analysis mainly focuses on the GPT-4 synthesized visual instruction tuning dataset. Recently, as the GPT-4V [3] becomes publicly available, there are some concurrent works (Wang et al., 2023a; Chen et al., 2023b) prompting GPT-4V with images as inputs to generate visual instruction tuning data. Future work can analyze the effect of tuning VLMs on such datasets and identify the advantages and disadvantages.

In our experiments, we mainly focus on the LLaVA-Architecture (Liu et al., 2023e) due to its strong performance and high efficiency. However, there are other foundation architectures such as Q-former in BLIP2 (Li et al., 2023d) and Perceiver Resampler in Flamingo (Alayrac et al., 2022). More diverse VLM architectures can be explored in the future to draw more general conclusions.

Acknowledgments

This research is based upon work supported by the U.S. DARPA ECOLE Program #HR001122S0052. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental

[3] https://openai.com/research/gpt-4v-system-card

purposes notwithstanding any copyright annotation therein.

References

Harsh Agrawal, Peter Anderson, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. nocaps: novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 8947–8956. IEEE.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a visual language model for few-shot learning.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Omri Avrahami, Dani Lischinski, and Ohad Fried. 2022. Blended diffusion for text-driven editing of natural images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18187–18197. IEEE.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023a. Qwen technical report. *CoRR*, abs/2309.16609.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. 2019. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 9448–9458.

Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. 2021. The mvtec anomaly detection dataset: A comprehensive real-world dataset for unsupervised anomaly detection. volume 129, pages 1038–1059.

Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie-Line Alberi-Morel. 2012. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. pages 1–10.

Ali Furkan Biten, Rubèn Tito, Andrés Mafla, Lluís Gómez i Bigorda, Marçal Rusiñol, C. V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas. 2019. Scene text visual question answering.

Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. 2015. HICO: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE International Conference on Computer Vision*.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023a. Shikra: Unleashing multimodal llm's referential dialogue magic. *CoRR*, abs/2306.15195.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023b. Sharegpt4v: Improving large multi-modal models with better captions. *CoRR*, abs/2311.12793.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023c. Can pre-trained vision and language models answer visual information-seeking questions? pages 14948–14968.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

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. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Chee-Kheng Chng, Chee Seng Chan, and Cheng-Lin Liu. 2020. Total-text: toward orientation robustness in scene text detection. *Int. J. Document Anal. Recognit.*, 23(1):31–52.

Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. 2009. NUS-WIDE:

a real-world web image database from national university of singapore. In *Proceedings of the 8th ACM International Conference on Image and Video Retrieval, CIVR 2009, Santorini Island, Greece, July 8-10, 2009*. ACM.

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. 2014. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Adam Coates, Andrew Y. Ng, and Honglak Lee. 2011. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, volume 15 of *JMLR Proceedings*, pages 215–223. JMLR.org.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *CoRR*, abs/2305.06500.

Luke Nicholas Darlow, Elliot J. Crowley, Antreas Antoniou, and Amos J. Storkey. 2018. CINIC-10 is not imagenet or CIFAR-10. *CoRR*, abs/1810.03505.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Stefan Lee, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2019. Visual dialog. volume 41, pages 1242–1256.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Mathias Eitz, James Hays, and Marc Alexa. 2012. How do humans sketch objects? *ACM Trans. Graph. (Proc. SIGGRAPH)*, 31(4):44:1–44:10.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. 2010. The pascal visual object classes (VOC) challenge.

Ali Farhadi, Ian Endres, Derek Hoiem, and David A. Forsyth. 2009. Describing objects by their attributes. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 1778–1785. IEEE Computer Society.

Li Fei-Fei, Rob Fergus, and Pietro Perona. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. page 178.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. 2017. Domain-adversarial training of neural networks. In Gabriela Csurka, editor, *Domain Adaptation in Computer Vision Applications*, Advances in Computer Vision and Pattern Recognition, pages 189–209. Springer.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. 2023. Llama-adapter V2: parameter-efficient visual instruction model. *CoRR*, abs/2304.15010.

Noa Garcia and George Vogiatzis. 2018. How to read paintings: Semantic art understanding with multi-modal retrieval. In *Computer Vision - ECCV 2018 Workshops - Munich, Germany, September 8-14, 2018, Proceedings, Part II*, volume 11130 of *Lecture Notes in Computer Science*, pages 676–691. Springer.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. 2019. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Gregory Griffin, Alex Holub, and Pietro Perona. 2007. Caltech-256 object category dataset.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. The false promise of imitating proprietary llms. *CoRR*, abs/2305.15717.

Jean-Philippe Thiran Guillaume Jaume, Hazim Kemal Ekenel. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *Accepted to ICDAR-OST*.

Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. Captioning images taken by people who are blind. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVII*, volume 12362 of *Lecture Notes in Computer Science*, pages 417–434. Springer.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. 2021a. The many faces of robustness: A critical analysis of out-of-distribution generalization. pages 8320–8329.

Dan Hendrycks and Thomas G. Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. 2021b. Natural adversarial examples. pages 15262–15271.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alexander Shepard, Hartwig Adam, Pietro Perona, and Serge J. Belongie. 2018. The inaturalist species classification and detection dataset.

Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge J. Belongie. 2015. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 595–604. IEEE Computer Society.

Qiang Hou, Weiqing Min, Jing Wang, Sujuan Hou, Yuanjie Zheng, and Shuqiang Jiang. 2021. Foodlogodet-1500: A dataset for large-scale food logo detection via multi-scale feature decoupling network. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4670–4679.

Ting-Yao Hsu, C. Lee Giles, and Ting-Hao Kenneth Huang. 2021. Scicap: Generating captions for scientific figures. pages 3258–3264.

Wenbo Hu, Yifan Xu, Yi Li, Weiyue Li, Zeyuan Chen, and Zhuowen Tu. 2023. BLIVA: A simple multimodal LLM for better handling of text-rich visual questions. *CoRR*, abs/2308.09936.

Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. 2007. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst.

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

EunJeong Hwang and Vered Shwartz. 2023. MemeCap: A dataset for captioning and interpreting memes.

Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert P. Dick, Hidenori Tanaka, Edward Grefenstette, Tim Rocktäschel, and David Scott Krueger. 2023. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1988–1997. IEEE Computer Society.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656.

Yannis Kalantidis, Lluis Garcia Pueyo, Michele Trevisiol, Roelof van Zwol, and Yannis Avrithis. 2011. Scalable triangulation-based logo recognition. In *Proceedings of the 1st International Conference on Multimedia Retrieval, ICMR 2011, Trento, Italy, April 18 - 20, 2011*, page 20. ACM.

Kimmo Kärkkäinen and Jungseock Joo. 2021. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 1547–1557. IEEE.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Min Joon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, volume 9908 of *Lecture Notes in Computer Science*, pages 235–251. Springer.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. 2011. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. volume abs/2305.01569.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2013, Sydney, Australia, December 1-8, 2013*, pages 554–561. IEEE Computer Society.

Elisa Kreiss, Fei Fang, Noah D. Goodman, and Christopher Potts. 2022. Concadia: Towards image-based text generation with a purpose.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images.

Anurendra Kumar, Keval Morabia, William Wang, Kevin Chang, and Alex Schwing. 2022. CoVA: Context-aware visual attention for webpage information extraction. In *Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pages 80–90, Dublin, Ireland. Association for Computational Linguistics.

Jason J Lau, Soumya Gayen, Dina Demner, and Asma Ben Abacha. 2019. Visual question answering in radiology (vqa-rad).

Yann LeCun. 1998. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*.

Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, José G. Moreno, and Jesús Lovón-Melgarejo. 2022. Viquae, a dataset for knowledge-based visual question answering about named entities. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 3108–3120. ACM.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2023a. MIMIC-IT: multi-modal in-context instruction tuning. *CoRR*, abs/2306.05425.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023b. Otter: A multi-modal model with in-context instruction tuning. *CoRR*, abs/2305.03726.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. 2017. Deeper, broader and artier domain generalization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5543–5551. IEEE Computer Society.

Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Hanwang Zhang, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, and Yueting Zhuang. 2023c. Fine-tuning multimodal llms to follow zero-shot demonstrative instructions.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023d. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. 202:19730–19742.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, Lingpeng Kong, and Qi Liu. 2023e. M^3it: A large-scale dataset towards multi-modal multilingual instruction tuning. *CoRR*, abs/2306.04387.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557.

Qing Li, Qingyi Tao, Shafiq R. Joty, Jianfei Cai, and Jiebo Luo. 2018. VQA-E: explaining, elaborating, and enhancing your answers for visual questions. 11211:570–586.

Shan Li and Weihong Deng. 2019. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1):356–370.

Yanda Li, Chi Zhang, Gang Yu, Zhibin Wang, Bin Fu, Guosheng Lin, Chunhua Shen, Ling Chen, and Yunchao Wei. 2023f. Stablellava: Enhanced visual instruction tuning with synthesized image-dialogue data. *CoRR*, abs/2308.10253.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023g. Evaluating object hallucination in large vision-language models. pages 292–305.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. 2023. Revisiting the role of language priors in vision-language models.

Krzysztof Lis, Krishna Kanth Nakka, Pascal Fua, and Mathieu Salzmann. 2019. Detecting the unexpected via image resynthesis. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2152–2161. IEEE.

Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2023a. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v(ision), llava-1.5, and other multi-modality models. *CoRR*, abs/2310.14566.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023b. Aligning large multi-modal model with robust instruction tuning. *CoRR*, abs/2306.14565.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023c. Mitigating hallucination in large multi-modal models via robust instruction tuning.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023d. Improved baselines with visual instruction tuning. *CoRR*, abs/2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023e. Visual instruction tuning.

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

Yuliang Liu, Lianwen Jin, Shuaitao Zhang, and Sheng Zhang. 2017. Detecting curve text in the wild: New dataset and new solution. *CoRR*, abs/1712.02170.

Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 1096–1104. IEEE Computer Society.

Yuen Peng Loh and Chee Seng Chan. 2019. Getting to know low-light images with the exclusively dark dataset. *Comput. Vis. Image Underst.*, 178:30–42.

Vincenzo Lomonaco and Davide Maltoni. 2017. Core50: a new dataset and benchmark for continuous object recognition. In *1st Annual Conference on Robot Learning, CoRL 2017, Mountain View, California, USA, November 13-15, 2017, Proceedings*, volume 78 of *Proceedings of Machine Learning Research*, pages 17–26. PMLR.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021a. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *The 59th Annual Meeting of the Association for Computational Linguistics (ACL)*.

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

Pan Lu, Liang Qiu, Jiaqi Chen, Tanglin Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021b. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Yadong Lu, Chunyuan Li, Haotian Liu, Jianwei Yang, Jianfeng Gao, and Yelong Shen. 2023. An empirical study of scaling instruct-tuned large multimodal models. *arXiv preprint arXiv:2309.09958*.

Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. 2023. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *CoRR*, abs/2306.09093.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. 2013. Fine-grained visual classification of aircraft. Technical report.

Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in Neural Information Processing Systems 27*, pages 1682–1690. Curran Associates, Inc.

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

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. 2022. Infographicvqa. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 2582–2591. IEEE.

Minesh Mathew, Dimosthenis Karatzas, R. Manmatha, and C. V. Jawahar. 2020. Docvqa: A dataset for VQA on document images. *CoRR*, abs/2007.00398.

Alexander Patrick Mathews, Lexing Xie, and Xuming He. 2016. Senticap: Generating image descriptions with sentiments. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 3574–3580. AAAI Press.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*, pages 1516–1525. IEEE.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*.

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating natural questions about an image.

Alex Olsen, Dmitry A. Konovalov, Bronson Philippa, Peter Ridd, Jake C. Wood, Jamie Johns, Wesley Banks, Benjamin Girgenti, Owen Kenny, James Whinney, Brendan Calvert, Mostafa Rahimi Azghadi, and Ronald D. White. 2018. Deepweeds: A multiclass weed species image dataset for deep learning. *CoRR*, abs/1810.05726.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. 2019. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1406–1415.

Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. 2017. Visda: The visual domain adaptation challenge. *CoRR*, abs/1710.06924.

Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2017. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IJCV*, 123(1):74–93.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. Connecting vision and language with localized narratives. In *ECCV*.

Jingyuan Qi, Minqian Liu, Ying Shen, Zhiyang Xu, and Lifu Huang. 2024. MULTISCRIPT: multimodal script learning for supporting open domain everyday tasks. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18888–18896. AAAI Press.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE.

Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. 2010. Adapting visual category models to new domains. In *Computer Vision - ECCV 2010, 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV*, volume 6314 of *Lecture Notes in Computer Science*, pages 213–226. Springer.

Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 2016. 300 faces in-the-wild challenge: database and results. *Image Vis. Comput.*, 47:3–18.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. 2019. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7374–7383.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: an open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-OKVQA: A benchmark for visual question answering using world knowledge. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VIII*, volume 13668 of *Lecture Notes in Computer Science*, pages 146–162. Springer.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. KVQA: knowledge-aware visual question answering. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 8876–8884. AAAI Press.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: A dataset for image captioning with reading comprehension. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part II*, volume 12347 of *Lecture Notes in Computer Science*, pages 742–758. Springer.

Vishwanath Sindagi, Rajeev Yasarla, and Vishal M. Patel. 2019. Pushing the frontiers of unconstrained crowd counting: New dataset and benchmark method. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 1221–1231. IEEE.

Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. 2021. Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 8802–8812. Computer Vision Foundation / IEEE.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. WIT: wikipedia-based image text dataset for multimodal multilingual machine learning. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 2443–2449. ACM.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: pre-training of generic visual-linguistic representations.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. EVA-CLIP: improved training techniques for CLIP at scale. *CoRR*, abs/2303.15389.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. pages 5100–5111.

Wei Ren Tan, Chee Seng Chan, Hernán E. Aguirre, and Kiyoshi Tanaka. 2019. Improved artgan for conditional synthesis of natural image and artwork. *IEEE Trans. Image Process.*, 28(1):394–409.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5228–5238. IEEE.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. 2017. Deep hashing network for unsupervised domain adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5385–5394. IEEE Computer Society.

Manisha Verma, Sudhakar Kumawat, Yuta Nakashima, and Shanmuganathan Raman. 2020. Yoga-82: A new dataset for fine-grained classification of human poses. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*, pages 4472–4479. Computer Vision Foundation / IEEE.

Oriol Vinyals, Charles Blundell, Timothy P. Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2016. Matching networks for one shot learning. *CoRR*, abs/1606.04080.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. 2011. Ucsd birds. Technical Report CNS-TR-2011-001, California Institute of Technology.

Haohan Wang, Songwei Ge, Zachary C. Lipton, and Eric P. Xing. 2019. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 10506–10518.

Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. 2023a. To see is to believe: Prompting GPT-4V for better visual instruction tuning. *CoRR*, abs/2311.07574.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 23318–23340. PMLR.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. 2023b. Image as a foreign language: BEIT pretraining for vision and vision-language tasks. pages 19175–19186.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners.

Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. 2017. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Trans. Geosci. Remote. Sens.*, 55(7):3965–3981.

Zhiyang Xu, Ying Shen, and Lifu Huang. 2023. Multi-Instruct: Improving multi-modal zero-shot learning via instruction tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11445–11465, Toronto, Canada. Association for Computational Linguistics.

Yue Yang, Artemis Panagopoulou, Qing Lyu, Li Zhang, Mark Yatskar, and Chris Callison-Burch. 2021. Visual goal-step inference using wikihow. pages 2167–2179.

Barry Yao, Sijia Wang, Yu Chen, Qifan Wang, Minqian Liu, Zhiyang Xu, Licheng Yu, and Lifu Huang. 2024. Ameli: Enhancing multimodal entity linking with fine-grained attributes. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2816–2834, St. Julian's, Malta. Association for Computational Linguistics.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 2733–2743. ACM.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. 2023. mplug-owl: Modularization empowers large language models with multimodality. *CoRR*, abs/2304.14178.

Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, Jing Shao, and Wanli Ouyang. 2023. LAMM: language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *CoRR*, abs/2306.06687.

Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. 2015. LSUN: construction of a large-scale image dataset using deep learning with humans in the loop. *CoRR*, abs/1506.03365.

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

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

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. 2023. Investigating the catastrophic forgetting in multimodal large language models. *CoRR*, abs/2309.10313.

Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. 2019. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *CoRR*, abs/2306.17107.

Bo Zhao, Yanwei Fu, Rui Liang, Jiahong Wu, Yonggang Wang, and Yizhou Wang. 2019. A large-scale attribute dataset for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0.

Bo Zhao, Boya Wu, and Tiejun Huang. 2023. SVIT: scaling up visual instruction tuning. *CoRR*, abs/2307.04087.

Bolei Zhou, Àgata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2018. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6):1452–1464.

Bolei Zhou, Àgata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. pages 487–495.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models.

Yutong Zhou and Nobutaka Shimada. 2021. Generative adversarial network for text-to-face synthesis and manipulation with pretrained BERT model. In *16th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2021, Jodhpur, India, December 15-18, 2021*, pages 1–8. IEEE.

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

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004.

A More Details on the Annotation Process of VISION-FLAN

A.1 Annotator Selection

Due to the complexity of the annotation task, we carefully design a selection process to select qualified annotators. Specifically, at beginning, the authors send out emails looking for graduate students in computer science who are interested in NLP and multi-modal learning. A group of 21 graduate computer science students signed up for a tutorial section. In the tutorial section, two PhD students in NLP explain the requirements for writing instructions, downloading the dataset and processing raw datasets into a unified format. After the tutorial, each candidate is assigned with three datasets and they have totally three days to process the raw datasets and write instructions. In the end, each candidate submits their annotations and two PhD students provide feedback to each candidate. The candidates then have two days to modify their instructions or formats based on the feedback. After two days, the candidates submit their final version of annotations and two PhD students discuss the quality of the annotations case by case. In the end, 7 out of 21 students were selected as qualified annotators. The compensation is 15$ per hour.

B Evaluation Datasets

We evaluate our models on several widely used multimodal evaluation benchmark datasets: (1) **MM-bench** (Liu et al., 2023f) is a comprehensive evaluation benchmark measuring VLM's capabilities from 20 dimensions. (2)**MME** (Fu et al., 2023) measures VLM's perception and cognition capabilities based on 14 diverse tasks. (3) **MM-Vet** (Yu et al., 2023) focuses on measuring the integration of various capabilities of VLMs, including OCR, recognition, knowledge, spatial awareness, math, and language generation. (4) **LLaVA-Bench** (Liu et al., 2023e) evaluates the instruction following and chat ability of VLMs in diverse daily-life visual tasks. (5) **POPE** (Li et al., 2023g) is an evaluation benchmark that probes object hallucination in VLMs. (6) **MMMU** (Yue et al., 2023) evaluates VLMs on multi-discipline tasks that require college-level subject knowledge and deliberate reasoning.

We also evaluate the newly proposed catastrophic forgetting problem (Zhai et al., 2023) of VLMs on 4 datasets: **CIFAR-10 and CIFAR-100** (Krizhevsky et al., 2009), **MNIST** (LeCun, 1998), and **miniImageNet** (Vinyals et al., 2016). We report the averaged performance of VLMs on the four benchmarks in the CF column of Table 2.

C Evaluation Protocols

For MM-Bench, MME, MM-Vet, LLaVA-Bench, POPE and MMMU, we use their official implementations of evaluation code[4] to evaluate the performance. Specifically, the evaluation scripts of MM-bench and MM-Vet call GPT-4 API to evaluate the correctness of a prediction given the target output and produce a binary score (0 or 1). Similarly, the evaluation of LLaVA-Bench also leverages GPT-4, and in addition to the target outputs, the evaluation method considers detail descriptions of images. The evaluation results are scores indicating not only the correctness but the human-preference of the predictions. MME and POPE are binary classification tasks and their evaluation is based on string matching between the predictions and target labels.

D Baselines

We compare our method with recent vision-language models. All the baselines listed below have similar architectures which consist of a pretrained LLM, a pretrained image encoder, and a bridging module that connects them. **BLIP-2** (Li et al., 2023d) utilizes the Q-Former to bridge a pretrained image encoder with a pretrained LLM, and achieves strong zero-shot capabilities. **InstructBLIP** (Dai et al., 2023) is a visual-instruction-tuned BLIP-2 (Li et al., 2023d) model. The instruction tuning dataset is a mixture of 13 academic datasets and the LLaVA (Liu et al., 2023e) dataset. **Shikra** (Chen et al., 2023a) focuses more on the object grounding capability and is instruction tuned on referential dialogue dataset and LLaVA dataset (Liu et al., 2023e), both of which are synthetically generated via GPT-4. **LLaVA** (Liu et al., 2023e) is the first VLM finetuned on GPT-4 synthesized visual instruction tuning dataset and achieves remarkable performance as a general-purpose visual chatbot. **Qwen-VL** and **Qwen-VL-Chat** (Bai et al., 2023b) are recently proposed VLMs based on Qwen (Bai et al., 2023a) language model and are trained on a large-scale (50 million instances) private visual instruction tuning dataset. **LLaVA-1.5** (Liu et al., 2023d) is a LLaVA model trained on a mixture of shareGPT [5], LLaVA (Liu et al., 2023e) and 8 academic image-text datasets.

E Additional Results

E.1 Effect of GPT-4 synthesized data on comprehensive evaluation benchmarks

We show additional results in Figure 8 to support our claim that training on GPT-4 synthesized data can not effectively improve VLMs' performance on comprehensive evaluation benchmarks. As we can observe, adding more training instances from the LLaVA dataset does not improve VISION-FLAN BASE's performance on MM-Bench.

E.2 Why VLMs Trained on VISION-FLAN are Better than State-of-the-Art VLMs?

In this section, we perform two case studies to explain why models trained on VISION-FLAN can perform better compared to state-of-the-art VLMs.

E.2.1 Case Study on OCR

When we manually check the predictions of VISION-FLAN CHAT and compare them to other VLMs, the first trend we observe is that VISION-FLAN CHAT can better perform OCR as shown in Figure 10. To quantify this observation, we evaluate LLaVA, LLaVA 1.5 and our models on the challenging TextOCR dataset (Singh et al., 2021). We ask the VLMs to predict all the text on each

[4] https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models/tree/Evaluation
https://mmbench.opencompass.org.cn/leaderboard
https://github.com/yuweihao/MM-Vet
https://github.com/haotian-liu/LLaVA/blob/main/docs/LLaVA_Bench.md
https://github.com/RUCAIBox/POPE
https://github.com/MMMU-Benchmark/MMMU

[5] https://sharegpt.com/

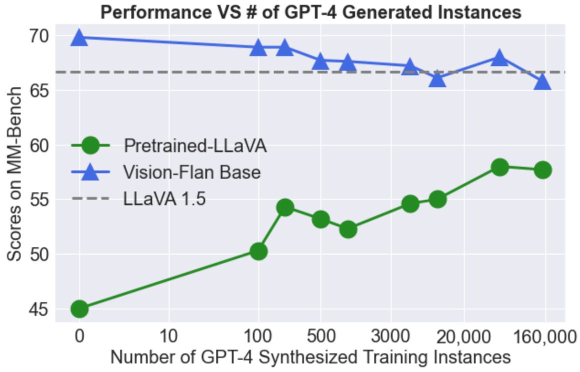


Figure 8: Effect of increasing the number of GPT-4 synthesized training instances on the comprehensive evaluation benchmark, namely MM-Bench. The dashed gray line indicates the performance of the-state-of-the-art LLaVA 1.5 model.

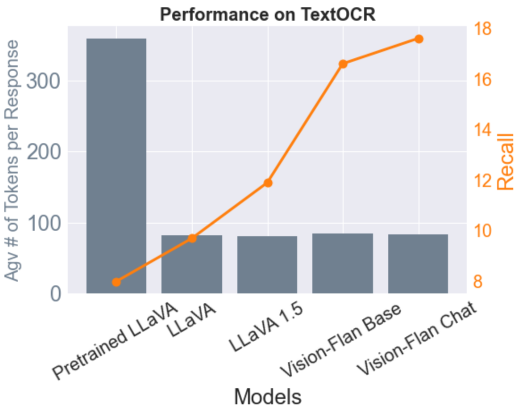


Figure 9: Performance of various VLMs on TextOCR. The gray bars shows the averaged number of tokens per prediction and the orange line show the recall of predictions.

image and check the overlap between the target list of text pieces and the predictions. As shown in Figure 9, the recall of VISION-FLAN BASE and VISION-FLAN CHAT is much higher compared to LLaVA 1.5 while the averaged numbers of predicted tokens per response are similar.

E.2.2 Case Study on Entity Recognition

We also spot that models trained on VISION-FLAN can better identify entities in an image while LLaVA 1.5 simply captions the appearance of entities in an image. A qualitative example is shown in Figure 11.

To compute quantitative results, we randomly sample 1,000 images with their captions from the WIT dataset (Srinivasan et al., 2021), in which the images are from Wikipedia pages and the captions commonly contain entities appearing in the images. We prompt VLMs to introduce the entities in the image with some background knowledge. We leverage the EntityRecognizer from spaCy [6] to recognize the entities in both predictions and ground truth captions and compute the percentage of target entities appearing in the predictions. As shown in Figure 12, it is clear that VISION-FLAN BASE and VISION-FLAN CHAT predict more entities in their responses (gray bars) and have higher coverage of entities (orange line) compared to LLaVA 1.5.

Visual Input Example: OCR

User	**Look at image and tell me what is the content. Your caption should involve the text in the image.**
LLaVA-1.5	**An advertisement for Telbruin, introducing their best call rates yet. The advertisement features a clock and several landmarks from around the world, including the Eiffel Tower, Big Ben, and the Space Needle. The advertisement claims that they are offering low rates as low as $0.25 per minute.**
Vision-Flan Chat	**The image features a red background and an advertisement for Telbru, a telecommunications company. The ad highlights the introduction of their best 098 call rates yet, which are as low as $0.25 per minute. The rates apply from 27th July 2014, and the ad encourages customers to visit their website or call their customer care for more information…**

Figure 10: An example from TextCap to show that Vision-Flan allows VLMs to better recognize text.

F Additional Analysis

F.1 The Bridging Module Can Be Shared Across LLMs with Same Architecture

Recent studies (Jain et al., 2023) in aligning and finetuning LLMs suggest that alignment happens very localized with pruning of a few weights or neurons to alter the style and format of outputs from LLMs, and does not substantially change the parameter space of LLMs. Following this finding, we hypothesize that *the MLP layers that map visual features into LLMs' embedding space*

[6] https://spacy.io/api/entityrecognizer

Model	MM-Bench	MME	LLaVA-Bench	Pope
Pretrained LLaVA-Architecture	45.0	936.3	32.4	51.9
+ LLaMA 2 Chat	45.3 (100.6)	557.0 (59.5)	59.2 (182.7)	66.9 (128.9)
VISION-FLAN BASE w/ frozen LLM	52.4	1107.3	41.6	83.3
+ LLaMA 2 Chat	46.6 (88.9)	1095.8 (99.0)	56.4 (135.6)	80.9 (97.1)
VISION-FLAN BASE	69.8	1537.8	38.5	85.9
+ LLaMA 2 Chat	47.2 (67.6)	852.6 (55.4)	69.9 (181.6)	66.1 (76.9)
VISION-FLAN CHAT	67.6	1490.6	78.3	86.1
+ LLaMA 2 Chat	47.0 (69.5)	869.6 (59.3)	74.6 (95.3)	65.8 (76.4)

Table 9: Results of replacing Vicuna 1.5 with LLaMA 2 Chat in four VLMs. The gray rows denote the performance of original models and blue rows denote the performance of the VLMs after replacing the LLMs. The number in each bracket denotes the percentage of VLMs' performance after integration of LLaMA 2 Chat, compared to their original performance.

Visual Input Example: Entity Recognition

User	**Can you give a short introduction to this painting?**
LLaVA-1.5	**The painting is a gold-toned artwork featuring a man and a woman embracing each other. The couple is depicted in a close and intimate manner, with the woman's head resting on the man's shoulder. The painting is rich in detail, with the couple's…**
Vision-Flan Chat	**The Kiss by Gustav Klimt is a painting that showcases the artist's unique and innovative style. It is one of the most recognizable paintings of the 20th century and is considered an icon of modern art. The painting depicts a couple in an embrace, their bodies forming a single, unified shape. They are surrounded by a golden aureole, which represents the sacred and the divine…**

Figure 11: An example from MM-Vet to show that Vision-Flan allows VLMs to better recognize entities.

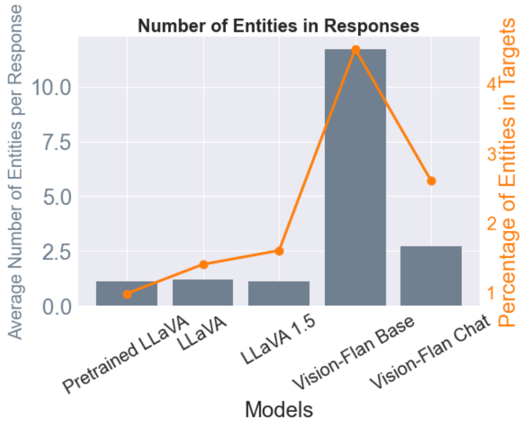


Figure 12: Performance of various VLMs on Entity Recognition. The gray bars show the average number of entities per response and the orange line shows the percentage of entities in the target response that appears in the prediction.

can be shared across LLMs with identical architecture but are tuned on different text alignment datasets. As shown in Table 9, we take four different models including VISION-FLAN BASE w/ frozen LLM which is finetuned on VISION-FLAN but with LLMs kept frozen as a case study, and directly replace their LLMs (Vicuna v1.5) with off-the-shelf LLaMA 2 Chat model. During inference, we use the official prompting template of LLaMA 2 chat instead of Vicuna v1.5. The results demonstrate that MLPs can be shared between LLMs with the same architecture but trained on different alignment datasets. An interesting observation is that there is a significant performance boost on LLaVA-Bench after we swap in LLaMA 2 Chat. If we finetune both the MLPs and the LLMs in VISION-FLAN BASE and VISION-FLAN CHAT, we observe a remarkable performance drop when we swap in LLaMA 2 chat. This is understandable because the LLaMA 2 chat can not effectively interpret visual features compared to the visual-instruction-tuned Vicuna v1.5.

F.2 Discrepancy Between Evaluation Benchmarks

In Table 2 and 9, we identify large performance discrepancy between multiple-choice benchmarks (e.g., MME and MM-Bench) and LLaVA-Bench on several models. Specifically, in Table 2, LLaVA achieves a score of 70.8 on LLaVA-Bench, comparable to the performance level of LLaVA 1.5. In contrast, LLaVA's performance on MME and MM-Bench is markedly lower, with scores of 1151.6 and

38.7, respectively, compared to LLaVA 1.5, which scores 1531.3 and 66.7. Furthermore, this trend is also evident in Table 9. Upon substituting the LLMs in VISION-FLAN BASE and VISION-FLAN CHAT with off-the-shelf LLaMA 2 Chat, both models exhibit a notable decline in performance on MME and MM-Bench, while maintaining comparable performance on LLaVA-Bench. Our hypothesis posits that LLaVA-Bench does not require LLM's strong understanding of the visual features, but rather relies on the language-prior of LLMs (Lin et al., 2023). Furthermore, the data synthesized by GPT-4 facilitates the model's ability to generate long-form responses, aligning with the preferences of the evaluation metric, namely, GPT-4 itself.

G Additional Related Work

Vision-Language Models. Previous works (Li et al., 2019; Chen et al., 2020; Tan and Bansal, 2019; Su et al., 2020; Wang et al., 2023b) mainly pretrain vision-language models (VLMs) from scratch with a unified masked-language modeling (MLM) objective (Devlin et al., 2019), which can impose significant training cost and inferior performance. Recently, a line of works proposes to build VLMs from the off-the-shelf visual encoders and LLMs by introducing a small amount of bridging parameters that maps visual features into the embedding space of LLMs. Flamingo (Alayrac et al., 2022) presents a VLM that is capable of processing interleaved image-text inputs and generating responses based on the visual content. It proposes Perceiver Resampler as the bridging module to connect the frozen LLM and visual encoder. OFA (Wang et al., 2022) proposes a sequence-to-sequence learning framework that maps images to discrete visual tokens and feeds the discrete visual tokens into LLMs. BLIP-2 (Li et al., 2023d) introduces Q-Former to bridge pre-trained and frozen vision and language models, based on which, MiniGPT-4 (Zhu et al., 2023) further adds a linear projector to bridge the gap between the visual encoder and language model encoder. LLaVA (Liu et al., 2023e) introduces a projector to fuse visual information into a large language model and unfreezes language model during visual instruction tuning.

H Datasets Used in VISION-FLAN

Dataset & Reference	Tasks
CINIC-10 (Darlow et al., 2018)	1. animal recognition in low resolution image 2. shipping method recognition in low resolution image 3. transportation option recognition in low resolution image 4. animal presence classification in low resolution image 5. object shipping object presence in low resolution image
MSCOCO (Lin et al., 2014)	1. multiple choice VQA 2. short image captioning 3. appliance recognition 4. furniture recognition 5. kitchen object recognition 6. vehicle recognition 7. animal recognition 8. sports object recognition 9. image text matching 10. image text selection
FairFace (Kärkkäinen and Joo, 2021)	1. human age classification 2. human gender classification 3. human race classification
IconQA (Lu et al., 2021b)	1. abstract diagram understanding 2. fill in blank in abstract diagram understanding
ImageNet-A (Hendrycks et al., 2021b)	1. object recognition of natural adversarial examples
ImageNet-C (Hendrycks and Dietterich, 2019)	1. blur type classification 2. coarse-grained image corruption classification 3. weather type classification 4. fine-grained image corruption classification
InfographicVQA (Mathew et al., 2022)	1. VQA 2. document level VQA
SemArt (Garcia and Vogiatzis, 2018)	1. painting time frame recognition 2. painting type recognition 3. painting school recognition 4. painting technique recognition 5. detailed image description
Set5 (Bevilacqua et al., 2012)	1. object recognition in low resolution image
TextCaps (Sidorov et al., 2020)	1. image captioning with reading comprehension
VisDial (Das et al., 2019)	1. visual dialogue with short context 2. visual dialogue with medium context 3. visual dialogue with long context 4. visual dialogue with very long context
STL-10 (Coates et al., 2011)	1. object recognition
Places365 (Zhou et al., 2018)	1. scene classification
Office-31 (Saenko et al., 2010)	1. image domain and office object classification 2. office object recognition

Dataset & Reference	Tasks
LSUN (Yu et al., 2015)	1. scene classification
FGVC-Aircraft (Maji et al., 2013)	1. aircraft family classification 2. aircraft manufacturer classification 3. aircraft variant classification 4. aircraft model classification
DeepFashion (Liu et al., 2016)	1. cloth texture classification
CUB-200-2011 (Wah et al., 2011)	1. bird species recognition
CLEVR (Johnson et al., 2017)	1. VQA in 3D rendered images 2. question answer matching 3. visual dialogue in 3D rendered images 4. VQA in 3D rendered images with multiple questions
CLEVR-CoGenT (Johnson et al., 2017)	1. VQA in 3D rendered images 2. question answer matching 3. VQA in 3D rendered images with multiple questions
A-OKVQA (Schwenk et al., 2022)	1. rationales generation 2. answer rationale generation 3. outside knowledge VQA
AI2D (Kembhavi et al., 2016)	1. diagram VQA
AID (Xia et al., 2017)	1. aerial scene classification
Caltech-256 (Griffin et al., 2007)	1. object recognition
CoVA (Kumar et al., 2022)	1. webpage recognition
DeepWeeds (Olsen et al., 2018)	1. weed species recognition
ExDark (Loh and Chan, 2019)	1. object recognition in low light environments
FFHQ-Text (Zhou and Shimada, 2021)	1. facial attribute textual descriptions generation
FlickrLogos-27 (Kalantidis et al., 2011)	1. logo recognition
FoodLogoDet-1500 (Hou et al., 2021)	1. food logo recognition
ImageNet-R (Hendrycks et al., 2021a)	1. object recognition in diverse image domain 2. image style classification
ImageNet-Sketch (Wang et al., 2019)	1. object recognition in sketch
JHU-CROWD++ (Sindagi et al., 2019)	1. scene classification
MNIST-M (Ganin et al., 2017)	1. number recognition
MVTecAD (Bergmann et al., 2021)	1. object anomaly detection 2. industrial item recognition

Dataset & Reference	Tasks
NABirds (Horn et al., 2015)	1. bird species recognition in north America 2. bird body parts detection
Road-Anomaly (Lis et al., 2019)	1. road anomaly detection
SCUT-CTW1500 (Liu et al., 2017)	1. curve text detection in the wild
Total-Text (Chng et al., 2020)	1. scene text detection and recognition
VisDA-2017 (Peng et al., 2017)	1. object recognition in 3D rendered image 2. multiple choice object recognition in 3D rendered image
Yoga-82 (Verma et al., 2020)	1. yoga pose recognition
Caltech101 (Fei-Fei et al., 2004)	1. object recognition 2. living organism classification
Cars (Krause et al., 2013)	1. car brand maker and year classification 2. car brand classification
Core50 (Lomonaco and Maltoni, 2017)	1. object recognition
NUS-WIDE (Chua et al., 2009)	1. animal presence classification
ObjectNet (Barbu et al., 2019)	1. object recognition
Places205 (Zhou et al., 2014)	1. indoor outdoor classification
300w (Sagonas et al., 2016)	1. indoor outdoor classification
Yahoo (Farhadi et al., 2009)	1. object recognition
LFW (Huang et al., 2007)	1. celebrity recognition
model-vs-human (Geirhos et al., 2019)	1. image-style classification
Office-Home (Venkateswara et al., 2017)	1. object recognition
Winoground (Thrush et al., 2022)	1. image caption matching
ConceptualCaptions (Sharma et al., 2018)	1. conceptual image captioning
KVQA+image question answer (Shah et al., 2019)	1. knowledge-aware VQA 2. visual entity recognition
MemeCap (Hwang and Shwartz, 2023)	1. meme understanding
PlotQA (Methani et al., 2020)	1. VQA over scientific plots
SentiCap (Mathews et al., 2016)	1. image captioning conditioned on sentiment
VQA-E (Li et al., 2018)	1. VQA 2. short image captioning
VQG (Mostafazadeh et al., 2016)	1. visual question generation 2. short image captioning

Dataset & Reference	Tasks
WIT (Srinivasan et al., 2021)	1. background knowledge extraction
WikiArt (Tan et al., 2019)	1. artist genre style recognition
VQA-RAD (Lau et al., 2019)	1. VQA in radiology
VOC2007 (Everingham et al., 2010)	1. multiple object recognition
VizWiz (Gurari et al., 2020)	1. answering visual questions from blind people 2. captioning image taken by blind people 3. quality issue classification of image taken by blind people
ViQuAE (Lerner et al., 2022)	1. knowledge based VQA about entities
ST-VQA (Biten et al., 2019)	1. scene text VQA
Stanford Dogs (Khosla et al., 2011)	1. dog species classification
Sketch (Eitz et al., 2012)	1. living organism classification in sketch 2. object recongnition in sketch
RAVEN (Zhang et al., 2019)	1. relational and analogical visual reasoning
PICKAPIC (Kirstain et al., 2023)	1. image prompt generation
PACS (Li et al., 2017)	1. object recognition in art painting 2. object recognition in cartoon 3. object recognition in photograph 4. dog image style classification 5. elephant image style classification 6. giraffe image style classification 7. guitar image style classification 8. horse image style classification 9. house image style classification 10. person image style classification
NOCAPS (Agrawal et al., 2019)	1. multiple short captions generation
Localized Narratives (Pont-Tuset et al., 2020)	1. COCO detailed image captioning 2. flickr30k detailed image captioning 3. open images detailed image captioning 4. ade20k detailed image captioning
INATURALIST (Horn et al., 2018)	1. class classification 2. family classification 3. genus classification 4. Latin English name classification 5. order classification 6. phylum classification 7. supercategory classification
HICO (Chao et al., 2015)	1. human activity detection
GEOMETRY3K (Lu et al., 2021a)	1. geometry question answering
FUNSD (Guillaume Jaume, 2019)	1. text detection in noisy scanned documents
FLICKR30K (Plummer et al., 2017)	1. multiple captions generation
DVQA (Kafle et al., 2018)	1. chart question answering
DTD (Cimpoi et al., 2014)	1. coarse grained texture classification 2. multiple texture detection

Dataset & Reference	Tasks
DOMAIN NET (Peng et al., 2019)	1. object recognition in clip art 2. object recognition in infograph 3. object recognition in painting 4. object recognition in quickdraw 5. object recognition in real image 6. image style classification
DOCVQA (Mathew et al., 2020)	1. document level VQA
DAQUAR (Malinowski and Fritz, 2014)	1. VQA
CONCADIA (Kreiss et al., 2022)	1. caption with background knowledge 2. short image captioning
Visual7W (Zhu et al., 2016)	1. VQA object attribute
VQAv2 (Goyal et al., 2017)	1. general VQA 2. question image matching
Visual Genome(Krishna et al., 2017)	1. spatial relationship question answering
OK-VQA(Marino et al., 2019)	1. outside knowledge VQA
ScienceQA (Lu et al., 2022)	1. VQA 2. explanation generation
OCR-VQA (Mishra et al., 2019)	1. VQA by reading text in image
wikiHow-image (Yang et al., 2021)	1. next step generation 2. image text step ordering 3. immediate next step selection 4. text image step ordering
SciCap (Hsu et al., 2021)	1. figure captioning
LAD (Zhao et al., 2019)	1. detailed object description generation
Dark Zurich (Sakaridis et al., 2019)	1. time of the day classification
RAF-DB (Li and Deng, 2019)	1. human emotion detection
GQA (Hudson and Manning, 2019)	1. spatial relationship question answering
VQA (Antol et al., 2015)	1. color 2. activity recognition 3. counting 4. object presence 5. object recognition 6. positional reasoning 7. scene recognition 8. sentiment understanding 9. sport recognition 10. utility affordance
Multimodal Factual Checking (Yao et al., 2023)	1. multimodal factual checking

I Task Categories in VISION-FLAN

Category	Tasks
Perception	1. CLEVR-CoGenT VQA in 3D rendered images 2. CLEVR-CoGenT question answer matching 3. CLEVR-CoGenT VQA in 3D rendered images with multiple questions 4. CLEVR VQA in 3D rendered images with multiple questions 5. GQA spatial relationship question answering 6. VQA color 7. VQA activity recognition 8. VQA counting 9. VQA object presence 10. VQA object recognition 11. VQA positional reasoning 12. VQA scene recognition 13. VQA sentiment understanding 14. VQA sport recognition 15. VQA utility affordance 16. VQA-E VQA 17. VQAv2 general VQA 18. Visual Genome spatial relationship question answering 19. CLEVR question answer matching 20. VizWiz answering visual questions from blind people 21. DAQUAR VQA 22. MSCOCO multiple choice VQA 23. Visual7W VQA object attribute 24. CLEVR VQA in 3D rendered images
Outside Knowledge	1. KVQA knowledge aware VQA 2. VIQUAE knowledge based VQA about entities 3. VQARAD VQA in radiology 4. OK-VQA outside knowledge VQA 5. A-OKVQA outside knowledge VQA
Reasoning	1. GEOMETRY3K geometry question answering 2. IconQA abstract diagram understanding 3. IconQA fill in blank in abstract diagram understanding 4. InfographicVQA VQA 5. InfographicVQA document level VQA 6. ScienceQA VQA 7. AI2D diagram VQA
OCR	1. DOCVQA document level VQA 2. DVQA chart question answering 3. PlotQA VQA over scientific plots 4. OCR-VQA VQA by reading text in image 5. ST-VQA scene text VQA

Category	Tasks
Document-Level OCR	1. FUNSD text detection in noisy scanned documents 2. SCUT-CTW1500 curve text detection in the wild 3. Total-Text scene text detection and recognition
Phrase-Level OCR	1. CoVA webpage recognition 2. FlickrLogos-27 logo recognition 3. FoodLogoDet-1500 food logo recognition
Knowledge Extraction	1. CONCADIA caption with background knowledge 2. KVQA visual entity recognition 3. WIT background knowledge extraction
Semantic Art Understanding	1. Semart painting time frame recognition 2. Semart painting type recognition 3. Semart painting school recognition 4. Semart painting technique recognition 5. Semart detailed image description 6. WikiArt artist genre style recognition
Visual Dialogue	1. CLEVR visual dialogue in 3D rendered images 2. Visdial visual dialogue with short context 3. Visdial visual dialogue with medium context 4. Visdial visual dialogue with long context 5. Visdial visual dialogue with very long context
Rational and Script Generation	1. ScienceQA explanation generation 2. A-OKVQA rationales generation 3. A-OKVQA answer rationale generation 4. MemeCap meme understanding 5. wikiHow-image next step generation 6. VQG visual question generation
Coarse-grained Captioning	1. ConceptualCaptions conceptual image captioning 2. FLICKR30K multiple captions generation 3. NOCAPS multiple short captions generation 4. PICKAPIC image prompt generation 5. VizWiz captioning image taken by blind people 6. VQA-E short image captioning 7. VQG short image captioning 8. MSCOCO short image captioning 9. CONCADIA short image captioning

Category	Tasks
Fine-grained Captioning	1. LAD detailed object description generation 2. FFHQ-Text facial attribute textual descriptions generation 3. Localized Narratives COCO detailed image captioning 4. Localized Narratives flickr30k detailed image captioning 5. Localized Narratives open images detailed image captioning 6. Localized Narratives ade20k detailed image captioning 7. SciCap figure captioning 8. SentiCap image captioning conditioned on sentiment 9. TextCaps image captioning with reading comprehension
Scene Classification	1. 300w indoor outdoor classification 2. AID aerial scene classification 3. Dark-Zurich time of the day classification 4. JHU-CROWD scene classification 5. LSUN scene classification 6. Places205 indoor outdoor classification 7. places365 scene classification
Animal Classification	1. CUB-200-2011 bird species recognition 2. DeepWeeds weed species recognition 3. INATURALIST class classification 4. INATURALIST family classification 5. INATURALIST genus classification 6. INATURALIST Latin English name classification 7. INATURALIST order classification 8. INATURALIST phylum classification 9. INATURALIST supercategory classification 10. NABirds bird species recognition in north America 11. NUS-WIDE animal presence classification 12. STANFORD DOGS dog species classification 13. NABirds bird body parts detection

Category	Tasks
Vehicle Classification	1. Cars car brand maker and year classification 2. Cars car brand classification 3. FGVC-Aircraft aircraft family classification 4. FGVC-Aircraft aircraft manufacturer classification 5. FGVC-Aircraft aircraft variant classification 6. FGVC-Aircraft aircraft model classification
Human Activity	1. HICO human activity detection 2. RAF-DB human emotion detection 3. Yoga-82 yoga pose recognition
Facial Recognition	1. LFW celebrity recognition 2. Fairface human age classification 3. Fairface human gender classification 4. Fairface human race classification
Anomaly Detection	1. Road-Anomaly road anomaly detection 2. MVTecAD object anomaly detection
General Object	1. Caltech-256 object recognition 2. Caltech101 object recognition 3. Caltech101 living organism classification 4. Core50 object recognition 5. ImageNet-A object recognition of natural adversarial examples 6. MNIST-M number recognition 7. MVTecAD industrial item recognition 8. ObjectNet object recognition 9. Office-Home object recognition 10. Office-31 image domain and office object classification 11. Office-31 office object recognition 12. STL-10 object recognition 13. Set5 object recognition in low resolution image 14. VOC2007 multiple object recognition 15. MSCOCO appliance recognition 16. MSCOCO furniture recognition 17. MSCOCO kitchen object recognition 18. MSCOCO vehicle recognition 19. MSCOCO animal recognition 20. MSCOCO sports object recognition 21. Yahoo object recognition

Category	Tasks
Complex Reasoning	1. RAVEN relational and analogical visual reasoning 2. Multimodal Factual Checking multimodal factual checking 3. wikiHow-image image text step ordering 4. wikiHow-image immediate next step selection 5. wikiHow-image text image step ordering
Image Text Matching	1. MSCOCO image text matching 2. Winoground image caption matching 3. MSCOCO image text selection 4. MSCOCO question image matching
General Object Classification in Special Image Domain	1. DOMAIN NET object recognition in clip art 2. DOMAIN NET object recognition in infograph 3. DOMAIN NET object recognition in painting 4. DOMAIN NET object recognition in quickdraw 5. DOMAIN NET object recognition in real image 6. ExDark object recognition in low light environments 7. ImageNet-R object recognition in diverse image domain 8. ImageNet-Sketch object recognition in sketch 9. PACS object recognition in art painting 10. PACS object recognition in cartoon 11. PACS object recognition in photograph 12. SKETCH living organism classification in sketch 13. SKETCH object recognition in sketch 14. Cinic-10 animal recognition in low resolution image 15. Cinic-10 shipping method recognition in low resolution image 16. Cinic-10 transportation option recognition in low resolution image 17. Cinic-10 animal presence classification in low resolution image 18. Cinic-10 object shipping object presence in low resolution image 19. VisDA-2017 object recognition in 3D rendered image 20. VisDA-2017 multiple choice object recognition in 3D rendered image

Category	Tasks
Image-Style Classification	1. DOMAIN-NET image style classification 2. ImageNet-R image style classification 3. PACS dog image style classification 4. PACS elephant image style classification 5. PACS giraffe image style classification 6. PACS guitar image style classification 7. PACS horse image style classification 8. PACS house image style classification 9. PACS person image style classification 10. Model-vs-human image style classification
Image Quality Classification	1. ImageNet-C blur type classification 2. ImageNet-C coarse-grained image corruption classification 3. ImageNet-C weather type classification 4. ImageNet-C fine-grained image corruption classification 5. VizWiz quality issue classification of image taken by blind people
Texture Classification	1. DTD coarse grained texture classification 2. DTD multiple texture detection 3. DeepFashion cloth texture classification

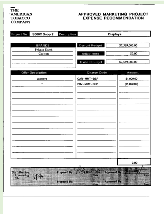
J VISION-FLAN Tasks

J.1 Generation Tasks

Task: Total-Text_scene_text_detection_and_recognition
Instruction: Detect and tell me all the text on the image. Please separate them with semicolons.
Target: Boots; inky; WINNER; G; TONY; AWARDSm; INCLUDING; BEST; MUSICAL; Winner; 4TONY; AWRDS; Best; Musical; An; American; Paris; Matilda

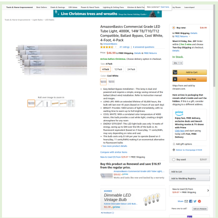
Task: SCUT-CTW1500_curve_text_detection_in_the_wild
Instruction: Detect all the text in the image.
Target: The texts are: Christmas; Happy.

Figure 13



Task: FUNSD_text_detection_in_noisy_scanned_documents
Instruction: Identify all the text in the image. Any ordering of the text is acceptable. Each chunk of text should be surrounded by double quotes and separated by commas.
Target: "Description", "Displays", "BRANDS", "Adjustment", "Carlton", "Amount", "Displays", "31,000.00", "(31,000.00)", "Distribution:", "", "5/5", "", "0.00", "Date", "Date", "Date", "Date", "THE AMERICAN TOBACCO COMPANY", "APPROVED MARKETING PROJECT EXPENSE RECOMMENDATION", "S0002 Supp 2", "Private Stock", "Current Budget", "$ 7,569,000.00", "Revised Budget", "$ 7,569,000.00", "$ 0.00", "CAR- MMT- DSP", "PRV- MMT- DSP", "Prepared By:", "Approved By:", "Approved By:", "Prepared By:", "", "Accounting File", "Project No", "Offer Description", "Charge Code"

Figure 14



Task: CoVA_webpage_recognition
Instruction: What is the name of the website based on the given image?
Target: The name of the website is www.amazon.com.

Task: FlickrLogos-27_logo_recognition
Instruction: Detect and provide the logo name in the image.
Target: The logo is Ford.

Task: FoodLogoDet-1500_food_logo_recognition
Instruction: What is the logo of the food product in the image?
Target: The logo in the image is Del Monte.

Figure 15

Task: CONCADIA_caption_with_background_knowledge
Instruction: Caption this image. Along with the caption, give me some background knowledge regarding this image.
Target: The beginning of the Regularis Concordia, drafted by thelwold in hermeneutic style. David Woodman gives a translation of the start of a charter drafted by "thelstan A", S 416 issued on 12 November 931:

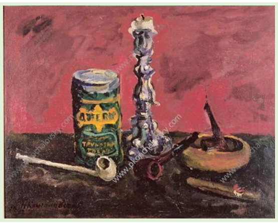
Task: KVQA_visual_entity_recognition
Instruction: You are provided with a picture. Write a caption for the image mentioning the name of the people present in the image and also their occasion depicted.
Target: Tryggve Gran in Antarctica (1911)

Task: WIT_background_knowledge_extraction
Instruction: Give me some background knowledge about this image. In your answer, you should include named entities.
Target: Alexandrovskaya Square is a city square of Taganrog.

Figure 16

Task: Semart_painting_timeframe_recognition
Instruction: Here is a picture of some fine art created in the past, I would like to know roughly what period in time it could have been made in?
Target: Roughly within the time 1601-1650

Task: Semart_painting_type_recognition
Instruction: Here is an image of some art. I want to know what type of painting it is. Among others, some types could be: religious, self-portraits, or landscapes
Target: This falls into the 'mythological' type



Task: WIKIART_artist_genre_style_recognition
Instruction: For the given painting, give me the artist name. Also provide the genre and style, if possible.
Target: This artwork was created by Pyotr Konchalovsky. It falls under the genre of still life and showcases Post Impressionism style of painting.

Figure 17

Task: Semart_painting_school_recognition
Instruction: I would like to know what school of art this painting could be from. Some possible schools of art are Italian, German, or Spanish among others
Target: This painting comes from the school of French art

Task: Semart_painting_technique_recognition
Instruction: Can you please tell me what technique is used to create the painting in this picture? Among some other techniques, it could be an oil painting?
Target: This painting uses Tempera and gold on paper, mounted on wood, 23 x 16 cm

Figure 18

Task: Semart_detailed_image_description
Instruction: Here is a picture of fine art, can you please give me a long description describing all of its details?
Target: Here is a description: The earliest reference to the painting appears in Francisco de los Santos, Descripción . . . del Escorial, published in 1657, when it was in the sacristy of the Pantheon: ' . . . commonly called the "Glory of El Greco" on account of the Glory above, but there is also represented, below, Purgatory and Hell, and on the other side, the Church Militant, with an immense number of the Faithful in adoration, raising their hands and eyes to Heaven, and among them Philip II; in the middle of the Glory is the Name of Jesus adored by Angels . . . signifying the words of Saint Paul In Nomine Jesu omne genu, flectatur Caelestium, Terrestrium, & Infernorum' (Saint Paul, Epistle to the Philippians, II, 10).The subject, then, is the Adoration of the Name of Jesus, a Jesuit counterpart of the Adoration of the Lamb, and incorporates the 'Church Militant', represented by the Holy League...

Figure 19

Task: Visdial_ visual_dialogue_with_medium_context
Instruction: This is a short conversation between two people, represented as A and B. They are talking about the given picture and are discussing some questions about it. Pay attention to the conversation and answer the question.
Input: A: what color is the counter? B: black tile A: can you see a microwave? B: no A: what color appliance? B: white A: is it a modern kitchen? B: i'd say so Question: is the sink stainless steel?
Target: no

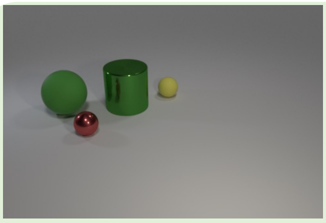
Task: Visdial_visual_dialogue_with_short_context
Instruction: Here is a short conversation between a human and an AI assistant. They are talking about a particular picture. They are discussing some questions and answers about it. Pay attention to the conversation and then answer the question at the end.
Input: Human: any people? Assistant: no Human: any animals? Assistant: no Question: is t it city?
Target: Assistant: no

Figure 20

Task: Visdial_visual_dialogue_with_long_context
Instruction: Here is a conversation between 2 people. They are talking about a picture and discussing some questions and answers about it. Read the conversation and then answer as the next person.
Input: Person1: how many motorcycles? Person2: 20 Person1: what color are they? Person2: black and white Person1: how many men? Person2: 20 Person1: is this in a city? Person2: no Person1: any women? Person2: no Person1: is a town? Person2: no Question: are they in leather?
Target: Person2: yes

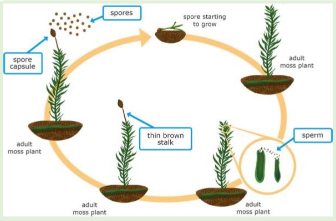
Task: Visdial_visual_dialogue_with_very_long_context
Instruction: Here is a detailed conversation between a person and a robot. They are discussing some questions and answers about a picture. From the conversation, answer the question as the robot.
Input: Human: what kind of uniform are the men wearing? Robot: they look like military Human: how many men would you guess there are? Robot: there are 3 men and a civilian woman Human: what color is the plane? Robot: white Human: what is the woman doing? Robot: talking to 1 of the men and pointing at the side of the plane Human: does the plane have writing on it? Robot: some numbers but i cannot make them out Human: does it look like a military plane? Robot: i think so but it is a close up, so hard to tell Human: are there any other planes in the image? Robot: yes i see parts of 2 others in the background Question: what is the woman wearing?
Target: Robot: a white and blue dress

Figure 21



Task: CLEVR_visual_dialogue_in_3D_rendered_images
Instruction: You are given some question and answer pairs as dialogue history, and you will answer the question at the end based on the image.
Input: What number of large blue matte things are the same shape as the big green matte object? 0 \n Is the number of red shiny objects left of the small red shiny object greater than the number of small gray metallic spheres? no \nDo the green thing to the left of the red object and the large thing that is right of the tiny red object have the same shape? no \nIs there any other thing that is the same size as the red metal object? yes \nWhat number of cubes are either big rubber things or tiny yellow rubber objects? 0 \nDoes the large sphere have the same material as the green cylinder? no \nThe large thing that is behind the matte ball in front of the matte thing that is on the right side of the large green metal cylinder is what shape?
Target: cylinder

Figure 22



Task: ScienceQA_explanation_generation
Instruction: Given a question and its answer, explain the rational behind the answer.
Input: Question: What can happen to the spores after they are released? Answer: They can grow into an adult moss plant.
Target: After they are released, spores can land on the ground and germinate. When spores germinate, they grow into adult moss plants. Unlike sperm and eggs, spores do not fuse.

Task: A-OKVQA_rationales_generation
Instruction: Provide 3 rationales for the given question and answer.
Input: The question is: What kind of precipitation is at the top of the mountain? The answer is: snow.
Target: The 3 rationales are: 1. The change in the atmosphere causes fog in higher elevations. 2. The precipitation is snow. 3. The mountain tops are white.

Task: A-OKVQA_answer_rationale_generation
Instruction: Answer the question and provide an explanation.
Input: The question is: What is the age of this woman?
Target: The answer is 60 because She is older with grey hair but not really old yet.

Figure 23

Task: MemeCap_meme_understanding
Instruction: Interpret and explain the meaning of the given Meme.
Target: Meme poster is trying to convey that don't like seeing their schoolmates outside school and will do anything to hide from them.

Task: Wikihow_next_step_generation
Instruction: The goal is to "Magazine Roll-Ups". Given current step specified by the content of the image and you have finished all previous steps: Zip up the sides of your boots in readiness for storing. Roll up an unwanted magazine and insert into one boot. Roll up a second unwanted magazine and insert it into the other boot. Put your boots in a standing up position in your shoe storage area.. What is the next step?
Target: Put your boots in a standing up position in your shoe storage area.

Task: VQG_visual_question_generation
Instruction: Given an image, generate a relevant question about the content of the image.
Target: What was being delivered?

Figure 24

Task: VQG_short_image_captioning
Instruction: Given an image, you will generate a very abstract caption for the image.
Target: Exterior of a brown house

Task: MSCOCO_short_image_captioning
Instruction: In this task, you are given an image and you will need to generate a short caption to describe it.
Target: A very cute stuffed animal with a candy bar.

Task: CONCADIA _short_image_captioning
Instruction: Can you describe the visual appearance of this image in one sentence?
Target: Red-and-white-striped lighthouse, behind a stand of palm trees

Figure 25

Task: PICKAPIC_image_prompt_generation
Instruction: Image you are writing a prompt for generating the given image. Please generate the prompt.
Target: Jedi duck holding a lightsaber.

Task: VQA-E+short_image_captioning
Instruction: You are provided with an image. Briefly Explain what is going on in the image.
Target: there are two people that are seated at the table talking

Task: VIZWIZ+captioning_image_taken_by_blind_people
Instruction: Give an image taken by blind people, generate multiple captions for this image.
Target: Caption 1: A woman's hand is resting of some kind of thick green bumpy fabric and her finger is under an orange food label Caption 2: Fingertip touching an orange juice carton that's sitting on top of dark knit fabric Caption 3: Orange juice is in the carton right at your fingertips. Caption 4: A person is touching a package of food.

Figure 26

Task: ConceptualCaptions_conceptual_image_captioning
Instruction: In this task, you will generate a short caption for the given image. The image could depict various scenes, objects, or people. Your goal is to provide a short and accurate caption that describes the main elements the image. Pay attention to the details, such as colors, objects, actions, and interactions between elements, to create a meaningful and informative caption.
Target: actor arrives at the premiere.

Task: FLICKR30K_multiple_captions_generation
Instruction: Each image has something going on. Carefully analyze the image and generate 5 short captions for each image.
Target: Caption 1: A young woman intently looking at something while another woman in a hat looks on. Caption 2: A young girl in glasses observes something in the distance. Caption 3: a teenager and her mom at a public place. Caption 4: A blond girl is fixing her sunglasses. Caption 5: A woman adjusts her sunglasses.

Figure 27

Task: VIZWIZ_captioning_image_taken_by_blind_people
Instruction: Give an image taken by blind people, generate multiple captions for this image.
Target: Caption 1: A woman's hand is resting of some kind of thick green bumpy fabric and her finger is under an orange food label Caption 2: Fingertip touching an orange juice carton that's sitting on top of dark knit fabric Caption 3: Orange juice is in the carton right at your fingertips. Caption 4: A person is touching a package of food.

Task: NOCAPS_multiple_short_captions_generation
Instruction: Give me a list of 10 short captions for this image.
Target: Caption 1: Group of people playing polo on bicycles on blacktop. Caption 2: A man riding a bicycle wearing a hat. Caption 3: Individuals playing a game consisting of riding on bicycles and hitting a ball with mallets. Caption 4: Three men in jackets playing croquet on bicycles. Caption 5: some people are playing hockey with their bicycle. Caption 6: Three people that are riding bicycles around on a basketball court. Caption 7: Several people ride bikes on what appears to be park asphalt. Caption 8: A group of people are riding the bicycle and playing hockey. Caption 9: Four people riding bicycles and playing polo while on a concrete pad near many houses. Caption 10: Four people playing polo on their bicycles on pavement.

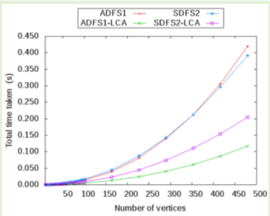
Figure 28

Task: LOC-NARRATIVES_COCO_detailed_image_captioning
Instruction: I want to know more about this image. Can you please describe it in detials?
Target: In this picture there are elephants at the right and left side of the image, there is water, grass land and rocks at the right and left side of the image, there are trees and dried grass at the background area of the image.

Task: LOC-NARRATIVES_flickr30k_detailed_image_captioning
Instruction: Please give me a detailed caption about the input image.
Target: In this image, there are two children in different color t-shirts, playing on an inflatable. One of these children is holding a toy and smiling. In the background, there are other toys, wooden objects and a person.

Task: LOC-NARRATIVES_open_images_detailed_image_captioning
Instruction: What is going in this image? Give me a detailed description.
Target: In this picture I can see food items in the jars with lids, which are on the wooden object, and there is blur background.

Figure 29



Task: SciCap+figure_captioning
Instruction: Generate a detailed caption for the given scientific figure from a paper. Your caption should mention the statistics in the figure and the language should be formal and clear.
Target: Comparison of total time taken and time taken by lca/la data structure by the most efficient algorithm for insertion of m = [BRACKET] edges for different values of n.

Task: SentiCap_image_captioning_conditioned_on_ sentiment
Instruction: You are provided with a picture, write a caption with a specific sentiment (positive or negative) related to the picture. Note that the sentiment in the caption should match the requested sentiment. Write a caption with a positive sentiment for the given image.
Target: a very nice looking tray filled with some excellent looking food

Task: textcaps_image_captioning_with_reading_comprehension
Instruction: Write a caption for the image. When you write the caption, also consider the text on the image and decide the best way to fit them into your caption.
Target: Here is a caption for this image: 'A blue Magic jersey with the number 50 sits against a grey background'

Figure 30

Task: LAD_detailed_object_description_generation
Instruction: I'll give you an image. What is the main object in it? Please describe the main object in great details.
Target: The object in the image is touring car. It is white, is big (compared to cars), weighs tons, is long, moves fast, has doors, has seats, has windows, has a engine, has a horn, has a steering wheel, has a brake, has a number plate, has four wheels, has lights, can move, can be driven, can carry a small number (≤10) of passengers, can carry a large quantity (> 1 tons) of goods, is for engineering, is for rescue, is for civil use, is for family, consumes diesel oil, can be used on urban roads, can be used on rural roads, is safe, is made of metal, is made of plastic, is expensive

Task: LOC-NARRATIVES_ade20k_detailed_image_captioning
Instruction: Can you give me a detailed description about this image?
Target: In this picture we can see few people and poles on the ground and we can see fences. In the background we can see a building with windows and we can see the name board.

Figure 31

Task: FFHQ-Text_facial_attribute_textual_descriptions_generation
Instruction: Generate some detailed descriptions for the face in the given image. Your description should contain the age, color of eyes, skin color and other facial attributes.
Target: This woman turns her face towards the front side. This girl is about 15 to 20 years old and has big brown eyes with double eyelids, bushy flat eyebrows and medium white to olive skin. A woman has bushy, medium flat eyebrows, brown eyes with double eyelids and a square face. This girl has a medium flat nose, a big natural mouth opened and a heart-shaped face. This woman is about 25 to 32 years old with long wavy brown hair cover up ears, square shape face and olive skin. This woman has small brown eyes with double eyelids, thick flat eyebrows, a big vegetable nose and a big natural mouth opened. A girl with a heart-shaped face, olive freckled skin and medium wavy cape blond hair cover up ears. A woman has medium wavy cape brown hair cover up ears and medium white to olive skin.

Figure 32

J.2 Classification Tasks

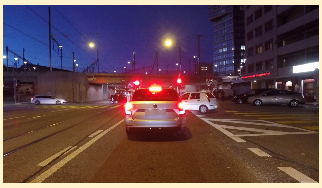
Task: LSUN_scene_classification
Instruction: In this task you will be provided with a picture of a scene (dining room, bedroom, kitchen, outdoor church, and so on) and you have to classify images into their corresponding scene categories. Your answer should be the name of the place. Options: (a) tower (b) classroom (c) dining room (d) bedroom (e) kitchen (f) church outside (g) living room (h) conference room (i) restaurant
Target: (h) conference room

Task: Places205_indoor_outdoor_classification
Instruction: In this task, you have to identify if the place or scene pictured is indoor or outdoor. In the image is among a total of 205 classes such as Hospital, Bridge, Courtyard, Motel,.... The classes of the images are a diverse set of places or scenes. Pay attention to the details as some of the images may contain an object that relates to a specific place while some images may directly show the place or scenary. So, your answer should be the place or scene shown in the image Options: (a) Outdoor (b) Indoor
Target: (b) Outdoor

Figure 33

Task: JHU-CROWD_scene_classification
Instruction: Provide the location of the scene in the image. It could be a water park, marathon, protest, stadium, or any other possible location.
Target: The scene is located at a stadium.

Task: AID+aerial_scene_classification
Instruction: You are given an aerial image. Tell me the scene in the image. The potential scenes are beach, industrial, meadow, and so on …
Target: The aerial scene is Airport.



Task: Dark-Zurich_time_of_the_day_classification
Instruction: Identify the time of the day when the image is captured. Options are: daytime, nighttime, twilight.
Target: The time of the day is twilight.

Figure 34

Task: places365_scene_classification
Instruction: Your task involves analyzing an image of a scene and identifying the appropriate name for that particular scene. Examples of scene names could include airfield, airplane cabin, airport terminal, alcove, alley, amphitheater, amusement arcade, etc.
Target: plaza

Task: 300w_indoor_outdoor_classification
Instruction: In this task, you will be presented with an image depicting a human portrait image. Your objective is to accurately classify the image by identifying the two categories it belongs to which are indoor and outdoor. To do so, carefully examine the visual elements present in the image, such as the background, people's clothes and any distinguishing features that can provide valuable clues for determining the category. For instance, if a person is at a baseball game outdoors, the category is outdoors. Once you have determined the category, provide your answer as the name of the category.
Target: Outdoor

Figure 35

Task: CUB-200-2011_bird_species_recognition
Instruction: Your objective is to identify the species of the bird depicted in the provided image.
Target: Long tailed Jaeger

Task: DeepWeeds_weed_species_recognition
Instruction: Identify weed species native to Australia in their natural habitat, alongside neighboring flora. **Target:** The weed species is Chinee apple.

Task: INATURALIST_class_classification
Instruction: Taxonomic category is a rank or group of organisms developed on the basis of their fundamental characteristics, similarities and dissimilarities. A class is a taxonomic rank above the order and below the phylum. Identify the class of the organism in the image.
Target: The class of the organism in the image is Magnoliopsida.

Figure 36

Task: INATURALIST_family_classification
Instruction: The family is a taxonomic rank above the genus and below the order. Identify the family of the organism in the image.
Target: The family of the organism in the image is Ranunculaceae.

Task: INATURALIST_genus_classification
Instruction: The genus is a taxonomic rank above the species and below the family. Identify the genus of the organism in the image.
Target: The genus of the organism in the image is Esox.

Task: INATURALIST_Latin_English_name_classification
Instruction: Identify the organism in the image. Give the english name(also called common name) followed by the scientific name(also called latin name). For example : "The organism in the image is Common Earthworm. Its scientific name is Lumbricus terrestris.
Target: The organism in the image is Blue-breasted Cordonbleu. Its scientific name is Uraeginthus angolensis.

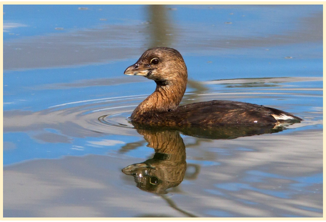
Figure 37

Task: INATURALIST_order_classification
Instruction: Taxonomic category is a rank or group of organisms developed on the basis of their fundamental characteristics, similarities and dissimilarities. The order is a taxonomic rank above the family and below the class. Identify the order of the organism in the image.
Target: The order of the organism in the image is Squamata.

Task: INATURALIST_phylum_classification
Instruction: Phylum is defined as a principal taxonomic category that ranks above class and below kingdom. Identify the phylum of the organism in the image.
Target: The phylum of the organism in the image is Tracheophyta.

Task: INATURALIST_supercategory_classification
Instruction: You will be given an image of an organism. Analyze the image and pick the super category for this organism from the options provided. Options: (a) Animalia (b) Reptiles (c) Insects (d) Ray-finned Fishes (e) Fungi (f) Amphibians (g) Birds (h) Plants (i) Mollusks (j) Mammals (k) Arachnids
Target: (c) Insects

Figure 38

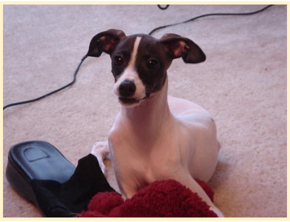


Task: NABirds_bird_species_recognition_in_north_America
Instruction: Identify the species of the bird in the image, considering its overall appearance, including size, shape, color, and patterns.
Target: Pied-billed Grebe

Task: NABirds_bird_body_parts_detection
Instruction: Tell me what body parts of the bird you can see in the image. provide the output in the following format: "The visible body parts are bill; crown; nape; left eye; belly; breast; back; tail; right wing.
Target: The visible body parts are bill; crown; nape; right eye; belly; breast; back; tail; left wing.

Task: NUS-WIDE_animal_presence_classification
Instruction: Identify if the given image contains any animal in it. Pay attention to each object in the image as well as the background environment while making this classification. If the image contains an animal, the answer should be 'yes'. Otherwise, 'no'. Options: (a) No (b) Yes
Target: (b) Yes

Figure 39



Task: STANFORD_DOGS_dog_species_classification
Instruction: Identify the breed of the dog in the image. Some sample classes are dhole, giant schnauzer, and leonberg.
Target: The dog breed is an Italian greyhound

Figure 40

Task: FGVC-Aircraft_aircraft_family_classification
Instruction: From the image provided to you, guess the family of the aircraft. Hint: Family: A family represents a collection of aircraft models produced by the same manufacturer, sharing common characteristics, design principles, or technological platforms.
Target: Spitfire

Task: FGVC-Aircraft_aircraft_ manufacturer_classification
Instruction: Determine the manufacturer of the provided aircraft image. The manufacturer refers to the company that designs, builds, and assembles the aircraft, possessing the expertise and experience in the aviation industry necessary for production and delivery.
Target: British Aerospace

Task: FGVC-Aircraft_aircraft_variant_classification
Instruction: Your objective is to analyze an aircraft image and provide the variant of the aircraft. (e.g., A300B4). Variant: A variant indicates a variation of a particular aircraft model, often incorporating specific modifications, improvements, or customizations compared to the base model.
Target: Yak-42

Figure 41

Task: FGVC-Aircraft_aircraft_model_classification
Instruction: Your objective is to analyze an aircraft image and provide the manufacturer, family, and variant of the aircraft in the specified order: manufacturer; family; variant (e.g., Airbus; A300; A300B4). Manufacturer: The manufacturer refers to the company that designs, builds, and assembles the aircraft, possessing the expertise and experience in the aviation industry necessary for production and delivery. Family: A family represents a collection of aircraft models produced by the same manufacturer, sharing common characteristics, design principles, or technological platforms. Variant: A variant indicates a variation of a particular aircraft model, often incorporating specific modifications, improvements, or customizations compared to the base model.
Target: Dornier; Dornier 328; Dornier 328

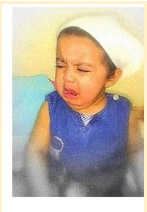
Task: Cars_car_brand_maker_and_year_classification
Instruction: In this task, based on the given image dataset of different cars, you have to identify the model + car make + Year of Make of a car in the image among a total of 196 categories such as Audi A5 Coupe 2012, BMW 3 Series Sedan 2012, Bentley Arnage Sedan 2009,... Pay attention to details such as the size, logo, type of the car to identify the model. So by looking at a car image, Give your answer in the following format: Model of the Car++Make of the Car++Year of Make
Target: GMC Savana Van 2012

Figure 42

Task: Cars_car_brand_classification
Instruction: In this task, you have to identify the brand of the car such as Audi, BMW, Bentley,... This means you have to identify the company which manufactured the car. For this, you need to look at the logo shown in the car image. Based on the detailing shown for the car image, the company model of the car can be identified. So, your answer should be the brand name of the car.
Target: Aston Martin

Figure 43

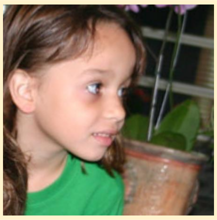
Task: HICO_human_activity_detection
Instruction: Answer a simple question. What is the person in the image doing? If there is no action being performed, describe the main object in the image.
Target: A person is skateboarding.



Task: RAF_DB_human_emotion_detection
Instruction: Give me details about the human in the image. What is their gender, race and age? What emotion are they depicting?
Target: The gender of the person is male. Their age range is 4-19 and their race is Asian. The emotion of the person in the image is Sadness.

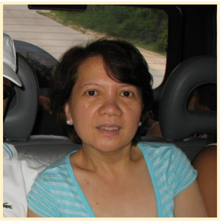
Task: Yoga-82_yoga_pose_recognition
Instruction: What is the name of the yoga pose?
Target: The yoga pose is Extended Revolved Triangle Pose or Utthita Trikonasana.

Figure 44



Task: Fairface_human_age_classification
Instruction: You are given an image of a person's face. This person can be of different ages, your task is to identify the person's age
Target: The person's age is 10-19

Task: Fairface_human_gender_classification
Instruction: Here is a picture of a person. Based only upon this picture, what would you guess this person's gender is?
Target: The person's gender is Female



Task: Fairface_human_race _classification
Instruction: What could be a good guess for this person's race in the given image?
Target: The person's race is Southeast Asian

Figure 45

Task: LFW_human_face_recognition
Instruction: In this task, you will be presented with a face image of an individual. Your objective is to accurately classify the image by identifying the person's identity it represents. To accomplish this, you must meticulously examine the facial features present in the image, such as the shape and structure of the face, eyes, nose, mouth, hair, and any other distinguishing features such as moles, scars, or birthmarks that can provide valuable clues for determining the identity. For instance, certain facial proportions, distinct eye color, or unique hair style could be defining characteristics of an individual's identity. Just as one might identify a bicycle by its wheels or a sunflower by its petals in other datasets, in this case, a person can be identified by their unique set of facial features. Once you've made an informed determination based on these visual clues, provide your answer as the identity of the person.
Target: Pete Sampras

Figure 46

Task: Road-Anomaly_road_anomaly_detection
Instruction: Detect the unusual dangers which can be encountered by a vehicle on the road.
Target: The dangers are lost tires.

Task: MVTecAD_object_anomaly_detection
Instruction: The primary objective of this task is to accurately identify the type and cause of anomalies in the object present in the provided image. The image depicts a specific category of object and texture, and within this category, there are defect-free images as well as images exhibiting different types of defects. Your task is to carefully examine the image and meticulously identify the specific type and cause of any deviations from the normal appearance of the object or texture. Pay close attention to irregularities in lines, shading, color scheme, and level of detail. Additionally, analyze the unique characteristics of the category, including shape, color, and texture. Your focus should be on precisely identifying the particular type and cause of the anomaly. The potential anomalies to consider encompass a wide range, such as gray strokes, bent objects, holes, missing wires, and more.
Target: The anomaly is combined.

Figure 47

Task: Caltech-256_object_recognition
Instruction: Your task is to identify the object category of a real-world image. The image can contain different objects like an American flag, bear, cake, and more. Analyze the shape, color, and texture of the object to determine its category. Consider the specific details of the label. Provide the name of the object based on your classification.
Target: coin

Task: Caltech101_object_recognition
Instruction: In this task, you have to classify the object in the image among classes such as Airplane, Ant, Butterfly, Chair,... The classes of the image are a diverse set ranging from objects to living beings. Pay attention to details as the object in the image can be in any format (sketch, painting, captured photo, etc) So, your answer should be the class of the object in the image
Target: garfield

Task: Caltech101_living_ organism_classification
Instruction: In this task, you have to classify if the setting contains a living thing or not. The object in the image is among a total of 102 classes such as Airplane, Ant, Butterfly, Chair,... The classes of the image are a diverse set ranging from objects to living beings. Pay attention to details as the object in the image can be in any format(sketch, painting, captured photo, etc) So, your answer should be if the object is a living thing or not. Options: (a) Yes (b) No
Target: (a) Yes

Figure 48

Task: Core50_object_recognition
Instruction: Your task is to identify the item shown in the picture. The images contain everyday objects such as a plug adapter, mobile phone, scissors, and more. It is important to carefully consider the object's shape, size, and color characteristics in order to accurately classify the image.
Target: cup

Task: Office-Home_object_recognition
Instruction: Your task involves classifying object images into their respective categories like Bed, Sink, Sneakers, Table, TV and so on; for instance, if the model is presented with an image of a laptop, it should correctly identify and categorize the image as 'Laptop'.
Target: Shelf

Task: MNIST-M_number_recognition
Instruction: In this task, you will be presented with a grayscale image containing a handwritten digit overlaid on a natural image background. Your objective is to correctly identify the digit in the image.
Target: 1

Figure 49

Task: ImageNet-A_object_recognition_of_natural_adversarial_examples
Instruction: In this task, given an image, please identify what the image contains a. The image could contain, among other things, animals, birds, daily objects, insects Options: (a) The provided image contains a lorikeet (b) The provided image contains a lion (c) The provided image contains an armadillo (d) The provided image contains a baseball player (e) The provided image contains a tricycle (f) The provided image contains a rugby ball (g) The provided image contains a jack-o'-lantern (h) The provided image contains a canoe
Target: (a) The provided image contains a lorikeet

Task: MVTecAD_industrial_item_recognition
Instruction: Your objective is to classify an image based on its corresponding object category. The image provided encompasses a diverse range of industrial items, including a bottle, cable, carpet, and more. Focus on the overall visual appearance of the image, paying attention to details such as lines, shading, color scheme, and level of detail. It is crucial to analyze the distinctive characteristics of the object, such as its shape, color, and texture, as these features may vary significantly between different object categories. Once you have completed the classification process, output the appropriate object name based on your analysis.
Target: The object is a pill.

Figure 50

Task: ObjectNet_object_recognition
Instruction: Your task is to recognize the object depicted in the given image. The object can be any item commonly used in our everyday lives, such as kitchen tools, food items, stationery, clothing, and more. To correctly identify the object, carefully observe its color, shape, and size characteristics.
Target: Flashlight

Task: Office-31_image_domain_and_office_object_classification
Instruction: The input for this task is an image of the commonly encountered office objects such as keyboards, file cabinets, and laptops from three different categories(AMAZON, DSLR, WEBCAM). The output of this task is to name the domain of image (AMAZON, DSLR, or WEBCAM) as well as the object in the image in the following format: 'Category Object Name' (exp. amazon mug) AMAZON: these images were captured from a website of online merchants, they are captured against clean background. DSLR cameras offer improved image quality when compared to standard webcams.
Target: webcam paper notebook

Figure 51

Task: Office-31_office_object_recognition
Instruction: The input for this task is an image of commonly encountered office objects such as keyboards, file cabinets, and laptops. from three different categories (Pictures from AMAZON, pictures taken with a DSLR camera, and pictures taken by WEBCAM). The output of this task is to name the object in the image.
Target: letter tray

Task: VOC2007_multiple_object_recognition
Instruction: Identify some objects that are present in the image. Give a comma separated list as the output.
Target: bottle, chair, tv monitor

Task: STL-10_object_recognition
Instruction: You will be presented with an image and your objective is to identify and name the object depicted in the image. Options: (a) airplane (b) horse (c) dog (d) bird (e) deer (f) truck (g) car (h) ship (i) monkey (j) cat **Target**: (a) airplane

Figure 52

Task: Set5_object_recognition_in_low_resolution_image
Instruction: In this task, recognize the subject in the image from among 5 subjects, namely - baby, bird, butterfly, head, woman.
Target: The subject in the image is a bird

Task: Yahoo_object_recognition
Instruction: In this task, you are given an image from a dataset, which contains images from different categories of animals, objects, and vehicles. These categories further divide into subcategories. Your job is to classify the given image into one of these subcategories, which could be anything from an aeroplane to a zebra. Your classification should be based on key identifiers like size, shape, color, distinctive features, and the context or environment depicted in the image. For example, if you're given an image of a zebra, your answer would simply be zebra. Remember that images could be of objects or vehicles as well. Your answer should be a single word representing the appropriate subcategory for the image, emphasizing specificity beyond the broad categories.
Target: building

Figure 53

Task: MSCOCO_appliance_recognition
Instruction: Given an image of a common electronic appliance from around the house, identify the type of object it is. It could be an appliance that is commonly used in the kitchen to cook or store food. Options: (a) This image contains an oven (b) This image contains a microwave (c) This image contains a toaster (d) This image contains a refrigerator (e) This image contains a sink
Target: (e) This image contains a sink

Task: MSCOCO_furniture_recognition
Instruction: Given an image of a piece of furniture in a house, identify the type of furniture. It is usually used to make the house look better and can be made of different kinds of material. Options: (a) This image contains a dining table (b) This image contains a bed (c) This image contains a toilet (d) This image contains a chair (e) This image contains a couch (f) This image contains a potted plant
Target: (d) This image contains a chair

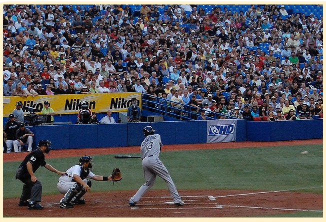
Figure 54

Task: MSCOCO_kitchen_object_recognition
Instruction: Given an image of something from the kitchen, identify what it could be. The image could be of cooking tools or items that are used for eating. It could also be used for serving food or storing it. Options: (a) This image contains a bottle (b) This image contains a cup (c) This image contains a wine glass (d) This image contains a fork (e) This image contains a knife (f) This image contains a bowl (g) This image contains a spoon
Target: (a) This image contains a bottle

Task: MSCOCO_vehicle_recognition
Instruction: Given an image of a vehicle, identify the kind of vehicle it is. The vehicle can be of different types; it could be something used, personal, or public transport. It could carry one or more people at the same time. Options: (a) This image contains a bus (b) This image contains a bicycle (c) This image contains a boat (d) This image contains an airplane (e) This image contains a motorcycle (f) This image contains a train (g) This image contains a truck (h) This image contains a car
Target: (b) This image contains a bicycle

Figure 55

Task: MSCOCO_animal_recognition
Instruction: Given an image of an animal, identify the kind of animal in the image. The picture could be of more popular animals that are visible around zoos or are sometimes domesticated at home. They could also sometimes be found in the wild. Options: (a) This image contains a cat (b) This image contains a dog (c) This image contains a cow (d) This image contains a bear (e) This image contains a sheep (f) This image contains a bird (g) This image contains an elephant (h) This image contains a zebra (i) This image contains a giraffe (j) This image contains a horse
Target: (h) This image contains a zebra

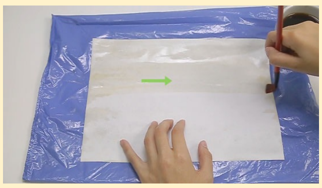


Task: MSCOCO_sports_object_recognition
Instruction: Given an image of sporting goods, identify what the object is. It could be used to play a team sport or an individual activity. The objects can also be used in different kinds of sports and sometimes make it easier for the wearer to play the sport. Options: (a) This image contains a ski (b) This image contains a surfboard (c) This image contains a frisbee (d) This image contains a baseball bat (e) This image contains a tennis racket (f) This image contains a baseball glove (g) This image contains a kite (h) This image contains a snowboard (i) This image contains a skateboard (j) This image contains a sports ball
Target: (j) This image contains a sports ball

Figure 56

Task: Wikihow_image_text_step_ordering
Instruction: You are doing Dipping Pine Cones in Paint. Is the step "Twist the end of a bamboo skewer into the top of the pine cone." the next or previous step to the step in the image? Options: (a) next (b) previous
Target: (b) previous

Task: Wikihow_immediate_next_step_selection
Instruction: You are doing Using an Oven to Dry Cilantro. What is the next step to step in the image? Options: (a) Store the dried cilantro leaves in an airtight container. (b) Preheat your oven to 250 °F (121 °C). (c) Remove the tray from the oven and let it cool for 10 minutes. (d) Spread the leaves on the baking tray to form 1 layer. (e) Wash the cilantro to remove dirt and debris.
Target: (c) Remove the tray from the oven and let it cool for 10 minutes.

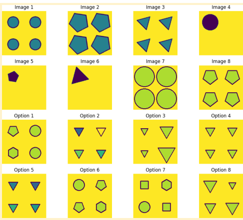


Task: Wikihow_text_image_step_ordering
Instruction: The goal is to "Thawing with a Microwave". Given the current step "Remove the plastic wrap and inspect your dough.", Is the picture the next or the previous step? Options: previous next
Target: next

Figure 57

Task: multimodal_factual_checking
Instruction: Context: Our Rating A widely-shared Facebook post claimed California had legalized 'pedophilia,' and that 'Now a 21 year old can have sex with an 11 year old, and not be listed on the sex registry as a sex offender.' That post and many like it are simply wrong. They grossly distort the proposals in state SB 145, which aims to eliminate a disparity in how LGBTQ young people are treated on California's sex offender registry. The legislation would eliminate automatic sex offeder registration for young adults who are convicted of having voluntary anal or oral sex with a minor and are within 10 years of age of the victim. Instead, a judge would make that decision, just as existing law allows judges to decide whether to place offenders in cases involving vaginal intercourse on the registry. The bill would not, in any fashion, make it legal for any adult to have any type of sex with a minor. The only change involves giving a judge discretion over whether to list an offender on the sex registry for certain sex acts. We rate the claims in the Facebook post Pants on Fire. PANTS ON FIRE - The statement is not accurate and makes a ridiculous claim. Does the context support "'PEDOPHILIA is now LEGAL in CALIFORNIA. Now a 21 year old can have sex with an 11 year old, and not be listed on the sex registry as a sex offender."? Options: (a) not sure (b) no (c) yes
Target: A1: (b) no

Figure 58



Task: RAVEN_relational_and_analogical_visual_reasoning
Instruction: Each image has 8 images labeled as Image 1 to Image 8. These 8 images follow a specific pattern. Detect the pattern and select the next image in the sequence from the 8 available options.
Target: Option 4

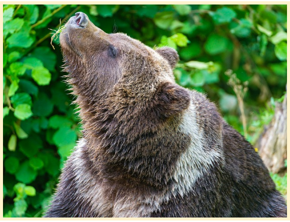
Figure 59

Task: image_text_matching
Instruction: Does "A woman in blue and purple holds a snowboard while standing in the snow." describes image? Options: (a) the description matches the image (b) the text is not a description of the image
Target: (a) the description matches the image

Task: Winoground+image_caption_matching
Instruction: In this task, you will be provided with an image and two captions. Your task is to identify which of the two captions correctly describes the image. Options: (a) the white wall will soon be painted blue (b) the blue wall will soon be painted white
Target: (a) the white wall will soon be painted blue

Task: image_text_selection
Instruction: Which option in the options that is the caption of the image. Options: (a) A couple of laptops with one sitting on a microwave. (b) Two older women are preparing for a dinner. (c) A desk with a computer monitor, printer and cd rack. (d) A girl preparing to put condiments on her dinner plate. (e) A man is taking an image on his phone of a bus.
Target: (d) A girl preparing to put condiments on her dinner plate.

Figure 60



Task: question_image_matching
Instruction: In this task, you need to decide if the image has enough information to answer "What does this man have hanging from his neck?" Options: (a) I can answer the question based on the image (b) I can not anser the question based on the image
Target: (b) I can not anser the question based on the image

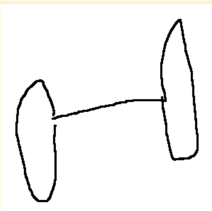
Figure 61

Task: DOMAIN-NET_object_recognition_in_clip_art
Instruction: Clip art is defined as simple pictures or symbols used in documents and presentations. The input is a clip art image. Identify the main object in the image.
Target: zigzag

Task: DOMAIN-NET_object_recognition_in_infograph
Instruction: An info graph is a visual image like a poster that is used to represent information or data about any object. For this task, the input will be a info graph. Identify the main object of the info graph.
Target: toaster

Task: DOMAIN-NET_object_recognition_in_painting
Instruction: The input for this task is a painting. Identify the main object in the painting.
Target: see saw

Figure 62



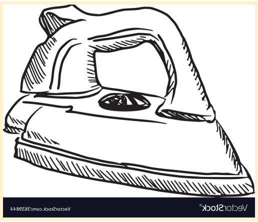
Task: DOMAIN-NET_object_recognition_in_quickdraw
Instruction: In this task, the input will be a rough sketch of something. Identify the main object depicted in the rough sketch.
Target: dumbbell

Task: DOMAIN-NET_object_recognition_in_real_image
Instruction: Identify the main object in the image.
Target: blueberry

Task: ExDark_object_recognition_in_low_light_environments
Instruction: The given image is taken in low-light environments. Identify the object in the image, including bicycle, boat, bottle, bus, car, and other objects.
Target: The object is Bicycle.

Figure 63

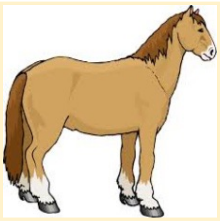
Task: ImageNet-R_object_recognition_in_diverse_image_domain
Instruction: Your task is to classify the image using various categories. You need to carefully observe the details of the object in the image, including its shape, color, and texture, as these characteristics may vary across different renditions. Output the appropriate object name as the result of your classification process.
Target: great white shark



Task: ImageNet_object_recognition_in_sketch
Instruction: You are given a sketch of an object. Tell me the name of the object in the image.
TargetThe sketch is a iron.

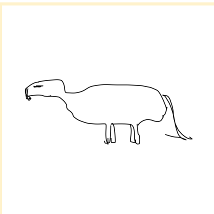
Task: PACS_object_recognition_in_art_painting
Instruction: You will be given an art painting image as input. Identify the main object in the image.
Target: dog

Figure 64



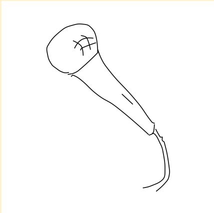
Task: PACS_object_recognition_in_cartoon
Instruction: You will be given an image of a cartoon. Identify the main object in the image.
Target: horse

Task: PACS_object_recognition_in_photograph
Instruction: The input is a photograph of an object. Identify the main object in the image.
Target: elephant



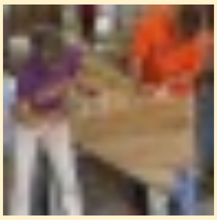
Task: SKETCH_living_organism_classification_in_sketch
Instruction: In this task, you will identify whether the picture contains a living organism. The images given are black and white sketches drawn by human beings. If the picture depicts a living organism or part of a living organism, the output should be "Living". Otherwise, print "Non-Living"
Target: Living

Figure 65



Task: SKETCH_object_recongnition_in_sketch
Instruction: Each image is a human drawn sketch of an object. Identify the main object in the image.
Target: microphone

Task: Cinic-10_animal_recognition_in_low_resolution_image
Instruction: The given image can contain various types of animals. Some of these animals are found in forests, drylands, or other natural areas. Some of them could also be domesticated pets. Please identify the animal in the picture.
Target: The image contains a bird



Task: Cinic-10_shipping_method_recognition_in_low_resolution_image
Instruction: The given image can contain different types of shipping equipment. They can carry goods across water or land, and they carry all types of materials required around the world. Please identify the type of shipping option in the picture.
Target: The image contains a ship

Figure 66

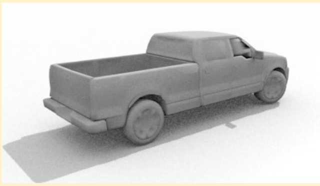
Task: Cinic-10_transportation_option_recognition_in_low_resolution_image
Instruction: The given image can contain different types of transport vehicles. People use these vehicles to travel around in their day-to-day lives. It could be air travel or a slower means of transport on the ground. Please identify the type of transport option in the picture.
Target: The image contains an automobile

Task: Cinic-10_animal_presence_classification_in_low_resolution_image
Instruction: The given image can contain some animals; they can be animals typically found in the wild or domesticated animals. The picture could also contain something that does not fit this description. Your job is to identify if the subject of the image is an animal or not.
Target: The object is an animal

Task: Cinic-10+object_shipping_object_presence_in_low_resolution_image
Instruction: The given image can contain some vehicles used for transporting goods and materials across large distances, even around the world. The picture could also contain something that does not fit this description. Your job is to identify if the subject of the image can be used for shipping goods or not.
Target: The object can be used for shipping

Figure 67

Task: VisDA-2017_object_recognition_in_3D_rendered_image
Instruction: Your task is to classify an image based on its corresponding object category. The image contains a variety of objects distributed among 12 categories, including aeroplane, horse, knife, person, plant, and others. To accurately classify the image, carefully analyze its visual characteristics, such as shape, color, texture, and spatial context relations, as these attributes can vary significantly across different domains. Once you have identified the object category of the image, output the appropriate label for your classification.
Target: plant



Task: VisDA-2017_multiple_choice_object_recognition_in_3D_rendered_image
Instruction: You are given an image which contains a 3D rendered object. Your goal is to identify the category of the object present in the image from the given options. Options: (a) knife (b) horse (c) train (d) bus (e) plant (f) skateboard (g) car (h) bicycle (i) truck (j) aeroplane
Target: (i) truck

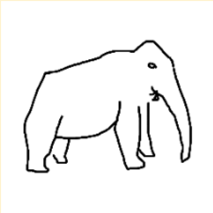
Figure 68

Task: DOMAIN-NET_image_style_classification
Instruction: You will be given an image. Answer 2 questions. What kind of image is this? Choose from clip art, info graph, painting, rough sketch, painting, real and sketch. Second question, what is the main object in the image? Answer it like "This is a clip art of an apple."
Target: This is a painting of a trumpet.

Task: ImageNet-R_image_style_classification
Instruction: Your goal is to classify the image based on its domain, which can be 'videogame', 'painting', 'sketch', 'cartoon', 'art', 'toy', 'deviantart', 'graphic', 'sculpture', 'misc', 'embroidery', 'sticker', 'graffiti', 'origami', or 'tattoo'. Your final output should specify the identified domain of the image.
Target: misc

Task: PACS_dog_image_style_classification
Instruction: You will be given an image of a dog. The image could be of different categories like painting, cartoon, photograph, or sketch. Identify the image category. Options: (a) Art painting (b) Cartoon (c) Sketch (d) Photograph
Target: (b) Cartoon

Figure 69



Task: PACS_elephant_image_style_classification
Instruction: You will be given an image of an elephant. The image could be of different categories like painting, cartoon, photograph, or sketch. Identify the image category.
Options: (a) Cartoon (b) Art painting (c) Photograph (d) Sketch
Target: (d) Sketch

Task: PACS_giraffe_image_style_classification
Instruction: You will be given an image of a guitar. The image could be of different categories like painting, cartoon, photograph, or sketch. Identify the image category.
Options: (a) Sketch (b) Cartoon (c) Art painting (d) Photograph
Target: (a) Sketch

Task: PACS_guitar_image_style_classification
Instruction: You will be given an image of a guitar. The image could be of different categories like painting, cartoon, photograph, or sketch. Identify the image category.
Options: (a) Cartoon (b) Photograph (c) Sketch (d) Art painting
Target: (b) Photograph

Figure 70

Task: PACS_horse_image_style_classification
Instruction: You will be given an image of a horse. The image could be of different categories like painting, cartoon, photograph, or sketch. Identify the image category.
Options: (a) Cartoon (b) Photograph (c) Art painting (d) Sketch
Target: (a) Cartoon

Task: PACS_house_image_style_classification
Instruction: You will be given an image of a house. The image could be of different categories like painting, cartoon, photograph, or sketch. Identify the image category.
Options: (a) Sketch (b) Photograph (c) Art painting (d) Cartoon
Target: (b) Photograph

Task: PACS_person_image_style_classification
Instruction: You will be given an image of a person. The image could be of different categories like painting, cartoon, photograph, or sketch. Identify the image category.
Target: Cartoon

Figure 71

Task: Model-vs-human_image_style_classification
Instruction: What is the artistic style of this image?
Target: power-equalisation

Figure 72

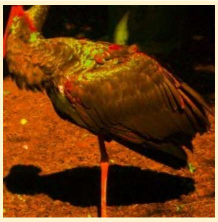
Task: ImageNet-C_blur_type_classification
Instruction: Given a blurred picture, identify the type of blur in the image, it can be blurred in different ways Options: (a) The image is corrupt, the specific corruption type is Glass blur (b) The image is corrupt, the specific corruption type is Defocus blur (c) The image is corrupt, the specific corruption type is Motion blur (d) The image is corrupt, the specific corruption type is Zoom blur
Target: (b) The image is corrupt, the specific corruption type is Defocus blur

Task: ImageNet-C_ coarse_grained_image_corruption_classification
Instruction: In this task, identify the type of corruption given a corrupted image. It could be digitally altered, contain natural distortions or contain other corruptions Options: (a) The corruption type is weather (b) The corruption type is blur (c) The corruption type is noise (d) The corruption type is digital
Target: (a) The corruption type is weather

Task: Vizwiz_quality_issue_classification_of_image_taken_by_blind_people
Instruction: Explain why the image quality is bad. Options: (a) rotation (b) bad framing (c) too bright (d) no flaws (e) blur (f) too dark (g) other (h) obscured
Target: (b) bad framing

Figure 73

Task: ImageNet-C_ weather_type_classification
Instruction: Given an image, identify what kind of weather conditions might have corrupted the image. It can be different types of bad weather or outdoor conditions Options: (a) The corruption type is snow (b) The corruption type is fog (c) The corruption type is frost (d) The corruption type is brightness
Target: (d) The corruption type is brightness



Task: ImageNet-C_fine_grained_image_corruption_classification
Instruction: Given an image, identify the type of corruption in the image. The image can have digitally generated noise, blur, or other distortions Options: (a) The corruption type is Saturate (b) The corruption type is Pixelate (c) The corruption type is Elastic transform (d) The corruption type is Contrast (e) The corruption type is Speckle noise (f) The corruption type is Shot noise (g) The corruption type is Gaussian blur (h) The corruption type is Spatter (i) The corruption type is Impulse noise (j) The corruption type is Gaussian noise
Target: (a) The corruption type is Saturate

Figure 74

Task: DTD+coarse_grained_texture_classification
Instruction: Texture is defined as the feel, appearance or consistency of a surface or substance from a human's perspective. Detect the primary texture represented in the image.
Target: cracked

Task: DeepFashion_cloth_texture_classification
Instruction: Can you write a very short description of the cloth?
Target: The cloth is an Abstract Mirrored Print Dress.

Task: DTD_multiple_texture_detection
Instruction: Texture is defined as the feel, appearance or consistency of a surface or substance from a human's perspective. Detect all the textures in the image. Present it as a comma separated list
Target: porous

Figure 75

J.3 VQA Tasks

Task: GQA_spatial_relationship_question_answering
Instruction: Answer the following question about the spatial relationship of objects in the given image. Your answer should be one or two words.
Input: The sign is on what?
Target: pole

Task: MSCOCO_multiple_choice_VQA
Instruction: Answer the given question by selecting an option.
Inputs: What is green on the plate? Options: (a) Salad. (b) Garnish. (c) Broccoli. (d) Tomato.
Target: (b) Garnish.

Task: VQA-E_VQA
Instruction: You are provided with an image and a question related to the image. Answer the question based on the information given in the image. Your answer should be a short phrase.
Input: How many players are there?
Target: 3

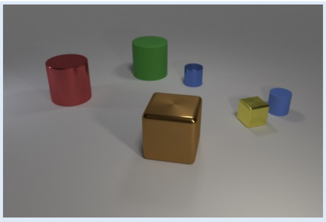
Figure 76

Task: VQA_color
Instruction: In this task, you are asked the color of some object in the image. Your answer should be a phrase.
Input: Question: What color is the umbrella primarily?
Target: orange

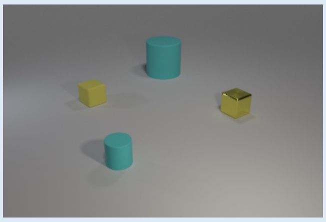
Task: Visual7W_VQA_object_attribute
Instruction: In this task, you will be asked about the attribute of some object. Your answer should be very concise.
Input: Question: What is the wall of the tub made of?
Target: tile

Task: VQA_activity_recognition
Instruction: In this task, you need to answer a question about the main activity happening in the image. Your answer should be one or two words.
Input: What is the girl doing?
Target: Eating.

Figure 77



Task: CLEVR_VQA_in_3D_rendered_images
Instruction: The input for this task is an image of 3D-rendered objects and a question that fall into different categories. The questions fall into five classes of tasks: Exist, Count, Compare Integer, Query Attribute, and Compare Attribute. The task here is to answer the question. Given me a very short answer.
Input: How many other metal cubes have the same color as the tiny shiny block?
Target: 0



Task: CLEVR-CoGenT_VQA_ in_3D_rendered_images_with_multiple_questions
Instruction: The input for this task is an image of 3D-rendered objects and a set of questions that fall into different categories. The questions fall into five classes of tasks: Exist, Count, Compare Integer, Query Attribute, and Compare Attribute. The output of this task is a set of answers to the given questions for each image. The answers should be generated based on the content of the image and the category of the question. The output should be in the form of text. The output should be in the format of "A1: YOUR ANSWER A2 YOUR ANSWER …"
Input: Q1: Are there any other things that are the same color as the metallic cube? Q2: The matte thing that is in front of the small yellow metal thing has what shape? Q3: What is the size of the cyan thing that is left of the cyan matte cylinder behind the yellow matte thing? Q4: Does the yellow thing that is in front of the tiny yellow rubber block have the same shape as the tiny thing that is in front of the tiny metal block? …
Target: A1: yes A2: cylinder A3: small A4: no A5: yes A6: 2 A7: large A8: no A9: rubber

Figure 78

Task: VQA_counting
Instruction: Please answer the question by counting the object mentioned in the question. Just output the number.
Input: Question: How many people can be seen?
Target: one

Task: VQA_object_presence
Instruction: Please answer the question about the presence of an object in the image. Given me a very brief answer.
Input: Are there any animals in the photo?
Target: yes

Task: VQA_object_recognition
Instruction: Answer a visual question about the class of the main object in the image. Given me a very abstract answer.
Input: What animal is this?
Target: elephant

Figure 79

Task: VQA_positional_reasoning
Instruction: In this task, the goal is to understand the location of objects within the presented image and provide an answer to the question. Given me a very short answer.
Input: What is to the right of cake?
Target: fork

Task: VQA_scene_recognition
Instruction: You are asked a question about the scene in the image. Answer the question with one or two words.
Input: Is this indoor or outdoor?
Target: indoor

Task: VQA_sentiment_understanding
Instruction: In this task, you will be asked a question regarding the emotion conveyed in the image. I need a short and concise answer.
Input: The question is Is this dog happy?
Target: yes

Figure 80

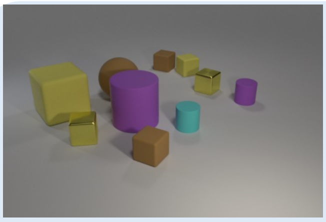
Task: VQA_sport_recognition
Instruction: Given a picture about sports, answer the following question. Answer the question with one or two words.
Input: What sport will the man be doing?
Target: surfing

Task: VQA_utility_affordance
Instruction: Please take a look at the picture and answer the following question by thinking about what each object in the picture can be used for. Your output can contain or two words.
Input: What is the man chopping?
Target: garlic

Task: VQAv2_general_VQA
Instruction: **Instruction**: Please take a look at the picture and answer a general question about the picture. Your output can contain one or two words.
Input: What color is the sign?
Target: red and white

Figure 81

Task: **Visual-Genome_spatial_relationship_question_answering**
Instruction: You are asked a question about the spatial relationship of objects in the image. Answer question with a short phrase.
Input: What is on the pizza?
Target: Ham

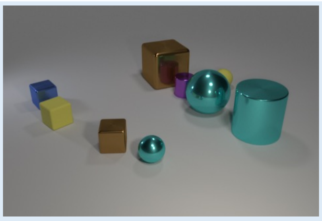


Task: CLEVR-CoGenT_question_answer_matching
Instruction: In this task, you will be presented with an image containing 3D-rendered objects along with a set of questions and corresponding answers. Your goal is to correctly match each question with its corresponding answer based on the visual content of the image. The output format should follow this pattern: Q1A3, Q2A5, Q3A2, Q4A1, Q5A1, indicating the question number followed by the corresponding answer number.
Input: Q1: How many other objects are there of the same color as the rubber ball? Q2: Is the color of the shiny object that is right of the cyan rubber cylinder the same as the big cylinder? Q3: What is the yellow object that is in front of the tiny cyan cylinder made of? Q4: Is the material of the large purple object the same as the large sphere? Q5: There is a yellow metal block that is behind the cyan rubber object; does it have the same size as the tiny cyan cylinder? ...
Target: Q1A5 Q2A7 Q3A8 Q4A6 Q5A6 Q6A3 Q7A2 Q8A1 Q9A4

Figure 82

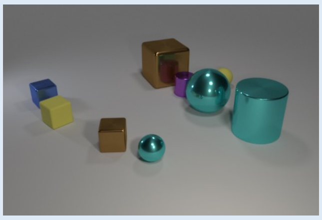
Task: Vizwiz_answering_visual_questions_from_blind_people
Instruction: A blind person asks you a question about this image, answer the question in the best way possible.
Input: What kind of food is this? Options: (a) canned beans (b) bushs reduced sodium dark red kidney beans (c) dark red kidney beans (d) kidney beans (e) reduced sodium kidney beans
Target: (b) bushs reduced sodium dark red kidney beans

Task: DAQUAR_VQA
Instruction: The input text will contain a question about the image. Answer the question. Your output should be one or two words.
Input: What is at the right side of the plant?
Target: cabinet



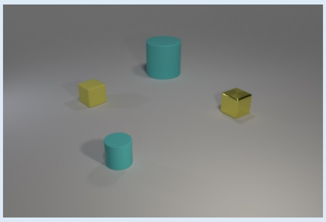
Task: **CLEVR-CoGenT_VQA_in_3D_rendered_images**
Instruction: The input for this task is an image of 3D-rendered objects and a question that fall into different categories. The questions fall into five classes of tasks: Exist, Count, Compare Integer, Query Attribute, and Compare Attribute. The task here is to answer the question and your answer should be one or two tokens.
Input: The cyan cylinder is what size?
Target: large

Figure 83



Task: CLEVR-question_answer_matching
Instruction: You will be given an Image of 3D-rendered objects, a number of Questions and same number of Answers. The task here is to match the questions to the right answers according to the image you see. The format of the output shoud be something like: Q1A3,Q2A5,Q3A2,Q4A1,Q5A1
Input: Q1: There is a shiny thing right of the big ball that is in front of the matte object in front of the matte cylinder; what is its shape? Q2: Is there any other thing that has the same size as the cube? Q3: Are there more balls that are behind the green metal sphere than rubber objects right of the gray matte ball? Q4: Is there a metal sphere on the right side of the rubber object on the left side of the yellow rubber thing?
Target: Q1A2 Q2A3 Q3A3 Q4A3 Q5A6 Q6A3 Q7A1 Q8A5 Q9A4

Figure 84

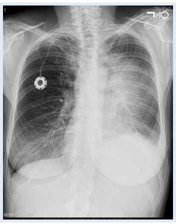


Task: CLEVR_ VQA_ in_3D_rendered_images_with_multiple_questions
Instruction: The input for this task is an image of 3D-rendered objects and a set of questions that fall into different categories. The questions fall into five classes of tasks: Exist, Count, Compare Integer, Query Attribute, and Compare Attribute. The output of this task is a set of answers to the given questions for each image. The answers should be generated based on the content of the image and the category of the question. The output should be in the form of text.
Input: Q1: Are there any other things that are the same color as the metallic cube? Q2: The matte thing that is in front of the small yellow metal thing has what shape? Q3: What is the size of the cyan thing that is left of the cyan matte cylinder behind the yellow matte thing? Q4: Does the yellow thing that is in front of the tiny yellow rubber block have the same shape as the tiny thing that is in front of the tiny metal block? …
Target: A1: yes A2: cylinder A3: small A4: no A5: yes A6: 2 A7: large A8: no A9: rubber

Figure 85

Task: KVQA_world_knowledge_enabled_VQA
Instruction: You are provided with a picture and a question related to the picture. Your job is to correctly answer the question with your background knowledge. Note that any references to directions (left, right, etc.) in the questions are from the perspective of the person depicted in the image. Your answer should consist of entity names.
Input: In which continent was the person in the image born?
Target: North America

Task: VIQUAE_ knowledge_based_VQA_about_entities
Instruction: With the help of this image, can you answer the question given in the input text by connecting the visual features in the image with named entities. Your answer should be a named entity.
Input: this mountain is the highest point in which country?
Target: Nam Chosun



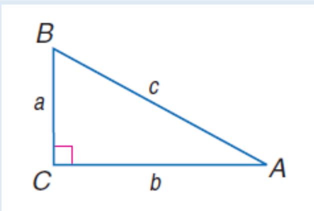
Task: VQARAD_VQA_in_radiology
Instruction: I will give you a radiology image(scan of a body part). Analyze it and answer the question given in the input text.
Input: Does the patient have a central line placed?
Target: Yes

Figure 86

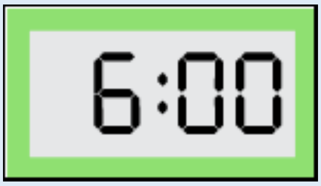
Task: OK-VQA_outside_knowledge_VQA
Instruction: Answer the following question about an image using your background knowledge outside of the given image. Your answer should be one or two words.
Input: What activity might these vehicles been used for?
Target: transportation

Task: A-OK-VQA _outside_knowledge_VQA
Instruction: Answer the question about the image. To correctly answer the question, you need think about knowledge outside the image. Your answer should be very short.
Input: What time period of the day is it?
Target: afternoon.

Figure 87



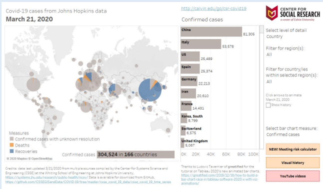
Task: GEOMETRY3K_geometry_question_answering
Instruction: I will give you a figure with some geometrical information. Analyze the image and data in the input text and answer the question.
Input: a = 8, b = 15, and c = 17, find \tan B. Options: (a) 2.43 (b) 1.88 (c) 1.67 (d) 1.23
Target: (b) 1.88



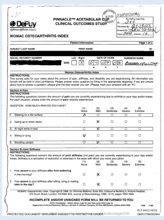
Task: Iconqa_abstract_diagram_understanding
Instruction: I have a question about the given abstract diagram, can you please give me a short answer?
Input: Ella is making her bed one morning. The clock shows the time. What time is it?
Target: The answer is 6:00 A.M.

Task: Iconqa_fill_in_blank_in_abstract_diagram_understanding
Instruction: Hey, here is an abstract diagram and sentence describing it. Can you help to fill in the missing part in the given sentence?
Input: The number _ is shown.
Target: 22

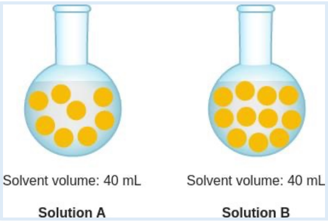
Figure 88



Task: Infographicvqa_VQA
Instruction: Here is an infographic representing some data in a visual form. Please answer my question using the data in the image.
Input: Which are the countries that have confirmed cases between 25000 and 50000?
Target: US, Spain

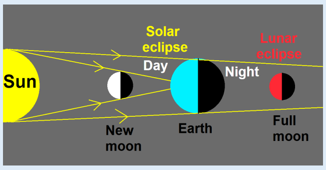


Task: Infographicvqa_document_level_VQA
Instruction: Here is a picture of a business or industry document, can you please answer my question based on the text in the document?
Input: What is the no of days given in the form?
Target: 358



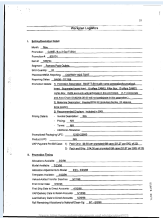
Task: ScienceQA_VQA
Instruction: You are asked a visual question within a science topic. The topics cover a diverse range of subject such as biology, physics, chemistry.
Input: Which solution has a higher concentration of yellow particles? Options: (a) neither; their concentrations are the same (b) Solution B (c) Solution A
Target: (b) Solution B

Figure 89

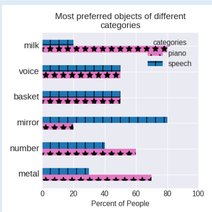


Task: AI2D_diagram_VQA
Instruction: Answer the multiple-choice question based on the diagram. The answer should be one of the choices. The question is:
Input: What provides the earth with solar energy? The choices are: (A) None of the above; (B) Sun; (C) New Moon; (D) Full Moon.
Target: The answer is: (B).

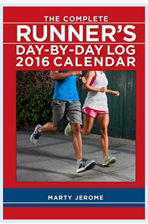
Figure 90



Task: DOCVQA_document_level_VQA
Instruction: Check the image and answer the question given in the input text.
Input: what is the reporting date mentioned ?
Target: 5/4/98 - 7/17/98



Task: DVQA_chart_question_answering
Instruction: I am trying to analyze this chart. Can you answer the question given in the input text?
Input: Is each bar a single solid color without patterns?
Target: no



Task: OCR-VQA_VQA_by_reading_text_in_image
Instruction: You are asked a question about the given image. Answer the question by reading the text written on the image.
Input: Which year's calendar is this?
Target: 2016

Figure 91