Guiding Vision-Language Model Selection for Visual Question-Answering Across Tasks, Domains, and Knowledge Types

Neelabh Sinha[1], Vinija Jain[2]*, Aman Chadha[3]†
[1]Georgia Institute of Technology   [2]Meta AI   [3]Amazon GenAI
nsinha68@gatech.edu, hi@vinija.ai, hi@aman.ai

Abstract

Visual Question-Answering (VQA) has become key to user experience, particularly after improved generalization capabilities of Vision-Language Models (VLMs). But evaluating VLMs for an application requirement using a standardized framework in practical settings is still challenging. This paper aims to solve that using an end-to-end framework. We present VQA360 – a novel dataset derived from established VQA benchmarks, annotated with task types, application domains, and knowledge types, for a comprehensive evaluation. We also introduce GoEval, a multimodal evaluation metric developed using GPT-4o, achieving a correlation factor of 56.71% with human judgments. Our experiments with state-of-the-art VLMs reveal that no single model excels universally, thus, making a right choice a key design decision. Proprietary models such as Gemini-1.5-Pro and GPT-4o-mini generally outperform others, but open-source models like InternVL-2-8B and CogVLM-2-Llama-3-19B also demonstrate competitive strengths, while providing additional advantages. Our framework can also be extended to other tasks[1].

1 Introduction

Visual Question Answering (VQA) (Antol et al., 2015) is the task of answering a question q about an image $\mathcal{I}$ correctly. This field has been faced with constant challenges in terms of the nature of the problem. For example, the question q can be about the image directly (Goyal et al., 2017a; Zhu et al., 2016; Goyal et al., 2017b), or outside the scope of the image with external knowledge (Marino et al., 2019; Schwenk et al., 2022). The images $\mathcal{I}$ can be a photograph, a mathematical chart (Masry et al.,
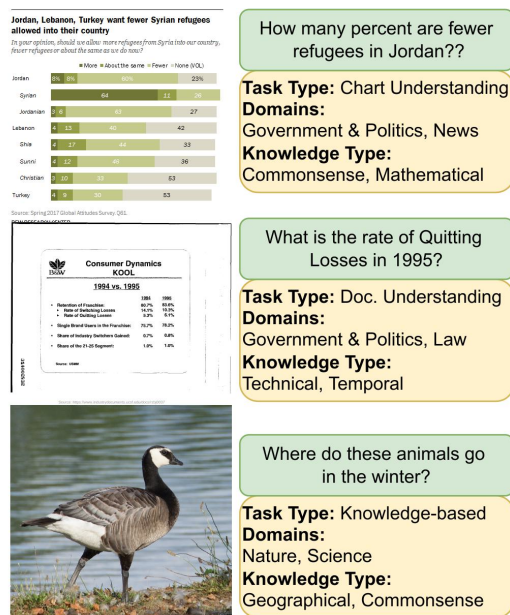

Figure 1: Examples of VQA360 tasks and their labels for task types, application domains, and knowledge type in our dataset.

2022; Li and Tajbakhsh, 2023), a document screenshot (Mathew et al., 2021), or more.

Dedicated methods (Zhu et al., 2016; Goyal et al., 2017b; Kafle and Kanan, 2017) have long existed to solve different challenges in VQA. But, with the advancement of Vision-Language models (VLMs) (Chen et al., 2023; Bai et al., 2023; Team et al., 2024; OpenAI et al., 2024; Liu et al., 2023a) in multimodal research, several applications have started adapting them due to their versatility. This is because after pre-training (Lin et al., 2024; Wei et al., 2024) on vast multimodal datasets (Chen et al., 2015; Thomee et al., 2016; Changpinyo et al., 2021; Masry et al., 2022; Mathew et al., 2021; Marino et al., 2019), VLMs can effectively generalize across various types of images, and can also incorporate external knowledge beyond the image.

But which VLM to utilize for a given VQA-based requirement? The complexity of this stems

---

*Work does not relate to position at Meta.

†Work does not relate to position at Amazon.

[1]Code and dataset can be found in the following link: https://github.com/neelabhsinha/vlm-selection-tasks-domains-knowledge-type.

from two directions - new VLMs being proposed and the diverse nature of tasks in VQA. New VLMs come up every month now (Liu et al., 2023a, 2024; Bai et al., 2023; team, 2024; Chen et al., 2023, 2024; Wang et al., 2023; Hong et al., 2024), and they differ in their architecture, training data, training strategy, size, etc., possessing different capabilites. In addition, users often face technical and business constraints in terms of compute, memory, cost of inference, data, and regulatory risks, which can favor specific VLMs over others. Second, tasks may vary from types such as Chart Question-Answering (Masry et al., 2022) or Document Understanding (Mathew et al., 2021), to application domains such as Science, or Sports, and the type of knowledge required, like Geographical Information, Mathematical Reasoning, and beyond. For an application that can fall into one or more such categories, how do you identify the best suited VLM? How to compare them meaningfully? These are gaps in existing literature. Technical reports of VLMs provide benchmarks and comparison, but they are very theoretical and limited.

To bridge the gap in evaluating VLMs on VQA, we propose an end-to-end framework that provides a standardized paradigm for evaluating vision-language models (VLMs) across three key aspects: *task type*, *application domain*, and *knowledge type*. Existing datasets like VQAv2 (Goyal et al., 2017a), OK-VQA (Marino et al., 2019), and ChartQA (Masry et al., 2022) offer task instances for training and evaluation but lack labels for other practical aspects. Our framework addresses this by developing and sharing a dataset VQA360 derived from the above benchmarks, where tasks are also labeled with their application domains and the knowledge types required for successful completion, as illustrated in Figure 1, allowing for a evaluation $360°$ in practical settings. Each task can have multiple tags for all these aspects. In addition, traditional NLP evaluation metrics have been shown to be poorly correlated with human judgments for generative models (Kamalloo et al., 2023; Liu et al., 2023b), an issue that extends to VLM. To address this, we introduce GoEval, a multimodal evaluation metric leveraging GPT-4o (OpenAI et al., 2024), which demonstrates superior alignment with human judgment compared to existing metrics. Together, they complement each other to provide an end-to-end, completely multimodal evaluation framework. Our framework evaluates 10 variations of 8 VLMs, accommodating diverse re-

quirements such as open-source, resource-efficient, or privacy-compliant models.

In summary, our aim is to address the following **research questions (RQs)**: **(1)** How to compare VLMs for different types of VQA tasks in practical settings? **(2)** How to evaluate those VLM outputs closely with human judgments? **(3)** As per current SOTA, which VLM is suited for which application, depending on various external constraints?

Our **key contributions** are as follows:

(i) We release VQA360 - a dataset of VQA tasks with three labeled aspects: *task types*, *application domains*, and *knowledge type*, enabling comparison based on different practical requirements.

(ii) We propose GoEval, which is a multimodal evaluation metric based on GPT-4o (OpenAI et al., 2024), and aligns more closely with human judgments for visual question-answering.

(iii) We analyze 10 variants of 8 state-of-the-art VLMs of different sizes and families, using our framework to compare their performance.

(iv) Using our analysis, we make recommendations on the best-suited VLMs for a given application requirement under different constraints.

2 VQA360: A Practical Evaluation Dataset

In this section, we discuss our dataset creation steps we followed in detail, which we propose to utilize for evaluating VLMs for VQA.

2.1 Source Datasets

To be able to evaluate VLMs in a wide variety of QA tasks, VQA360 is created from five standard datasets - VQAv2 (Goyal et al., 2017a), OK-VQA (Marino et al., 2019), A-OKVQA (Schwenk et al., 2022), ChartQA (Masry et al., 2022) and DocumentVQA (Mathew et al., 2021). VQAv2 is an extensive VQA benchmark, while OK-VQA and A-OKVQA are used primarily for knowledge-based VQA, where the answer to the question does not lie within the scope of the image. ChartQA and DocumentVQA are taken to evaluate the performance of VLMs on questions based on mathematical graphs and charts, and documents, respectively. From the test split of each dataset $\mathcal{D}_{test}$, we take $\max(|\mathcal{D}_{test}|, 1145)$ task instances, randomly sampled without replacement. 1145 was used as it is the minimum size of test set among the five datasets. Thus, our final experimental set contains 5725 task instances, with equal contributions from each dataset.

Aspect	Considered Tags
Application Domains	Anthropology, Formal logic, Economics, History, Law, Government and Politics, Linguistics, Computer Science, Mathematics, Science, Books, Fiction, Movies, News, Reviews, Justice, Professions, Public Places, Knowledge Base, Nature, Nutrition and Food, Social Media, Sports
Knowledge Types	Commonsense Knowledge, Visual Knowledge, Cultural Knowledge, Geographical Knowledge, Temporal Knowledge, Social Knowledge, Scientific Knowledge, Technical Knowledge, Mathematical Knowledge, Literary Knowledge, Other

Table 1: Application domains and knowledge types considered to label all the task instances.

2.2 Instance Tags

Using the dataset mentioned above only allows us to differentiate based on task types. But there are more ways in which a practical application can be classified. To enable that, we also classify VQA360 in two more aspects - application domains, and knowledge types. This is inspired from a recent work (Sinha et al., 2024), but adapted to suit this task. The application domain is the field a task belongs to, such as history, science, sports, etc., and the knowledge type is a specific type of knowledge required, such as geographic, common sense, etc.

Our initial set of tags is crafted manually, with an aim to achieve a broad spectrum of application domains and reasoning types. They are specified in Table 1, and cover a wide range of domains and knowledge types. We tag each of our task instances in the dataset with one or more application domains and knowledge types using the method described in Figure 2 and discussed in the following.

2.3 Generating Instance Tags

For creating the tags for task types, we map ChartQA to 'Chart Understanding', DocumentVQA to 'Document Understanding', A-OKVQA and OKVQA to 'Knowledge-based Visual Question-Answering' and VQAv2 to 'Visual Question-Answering'.

For application domains and knowledge types, we use gpt-3.5-turbo (Brown et al., 2020; OpenAI, 2023). To correctly generate instance tags, we require key features of the image and question, and also need to eliminate less useful information from the image to avoid confusions. To achieve this, following a recent work (Fu et al., 2023), we generate captions of images from VIT-GPT2[2], and object tags from the Azure Computer Vision API. Using these two as the descriptors of the image and the question, we prompt gpt-3.5-turbo to get the application domains and knowledge type.

The prompt used for this task is given in Table 7 in Appendix A.

After this, we post-process the tags by removing entries that do not belong to any of the entries in Table 1. If all tags are removed for a task instance, we add "Other" by default. Finally, we manually go through each of the labels and ensure that they are correctly tagged and replace any erroneous tag. The final set contains questions, images, candidate answers, task type, application domains, and knowledge type of all 5725 candidates.

From the instance tags, we remove the tags for which number of instances are less than 300 from our analyses. Please note, we do not remove the task instances entirely, as they may contain other tags included in the study, but just not consider those tags in reporting our results in Section 4 due to less number of instances. This gives a final set of 5 task types, 14 application domains, and 9 knowledge types, which are shown in Figure 3. We also report statistis of VQA360 in Table 2.

Statistic (per instance)	Mean	Std.	Max
Caption Length	46.56	8.84	96
Object Tags	13.01	7.83	59
Application Domains	1.7	0.68	7
Knowledge Type	2.19	0.83	9

Table 2: Average, std. and maximum of length of caption and count of object tags in generated image descriptors, and number of application domains and knowledge types per instance. This clarifies that significant number of task instances have multiple application domains and knowledge type tags.

VQA360 allows extended analysis of diverse VQA tasks and allows looking into performance of VLMs from entirely new perspectives. Further, the creation methodology can also be extended for enriching other datasets and creating benchmarks for evaluating VLMs in different settings.

We use VQA360 for rest of the analysis, and also release it publicly (linked in footnote of Page 1), for the research community to utilize in future research.

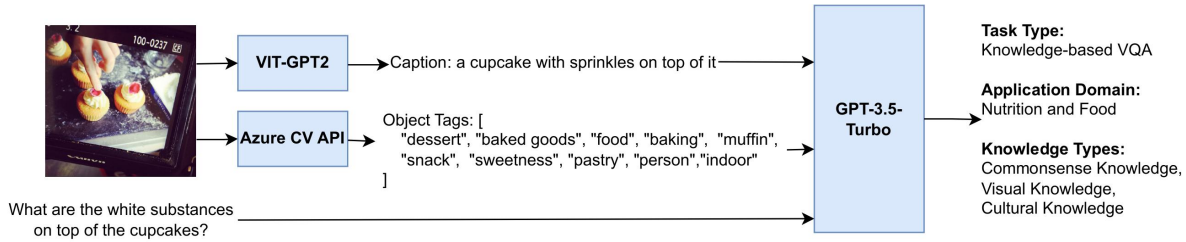[2]https://huggingface.co/nlpconnect/vit-gpt2-image-captioning

Figure 2: An example with steps taken to generate the instance tags for application domains and knowledge type (task type is mapped directly from the dataset the image is taken).
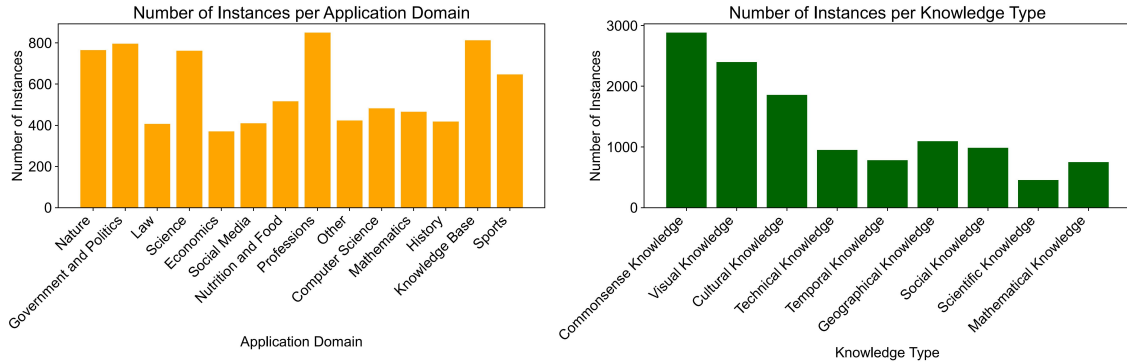

Figure 3: Number of task instances per application domain (left) and knowledge type (right) after generating the instance tags using GPT-3.5-Turbo. All categories are represented by approx 400 instances, which is sufficient for a representative analysis. Categories with < 300 instances are filtered out, and a task instance can be tagged to multiple categories of a single aspect.

Although it contains instances of existing benchmarks, it is a one-stop benchmark for an extended evaluation with labels for task types, application domains, and knowledge types. We also provide object tags and captions.

3   GoEval: A VQA Evaluation Metric

Evaluating QA using lexical matching has significant limitations, particularly when correct answers don't match with set of gold answers (Kamalloo et al., 2023). A recent work (Kamalloo et al., 2023) evaluated traditional VQA metrics against GPT-based evaluation for open-domain QA, and found it to be more aligned with human judgments.

Another alternative is to evaluate using NLG metrics such as ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and then use a threshold to determine correct/wrong. However, they also suffer from similar limitations, as evaluated from previous work (Liu et al., 2023b). A probable solution is to use BERTScore (Zhang et al., 2019), which compares texts in embedding space, focusing more on semantic similarity.

These metrics may be promising, but they are not equipped to handle multimodal settings. To fill this gap, we propose GoEval - a multimodal metric based on GPT-4o, which can be used to evaluate VQA. Similar to existing works (Kamalloo et al., 2023; Liu et al., 2023b), we create a prompt, and ask GPT-4o if the generated answer is correct. However, to incorporate the vision modality in making judgments, we also use the image. We use zero-shot evaluation with prompting, using both GPT-4o and GPT-4o-mini, to compare and contrast performance v/s cost trade-offs.

Formally, we pick a prompt function $\mathcal{P}$ from the first row of Table 8 of the Appendix A, and generate a prompt $p = \mathcal{P}(q, r, c)$ based on question q, the reference answer set r and the candidate response c. Using this prompt and the image, we prompt GPT-4o (OpenAI et al., 2024; OpenAI, 2024) to ask if it is correct. We also compare our prompt against two other ways – without using the image, and without reference answers. This is to understand which technique shows highest correlation with human judgments. We don't compare with traditional methods, as it is already done in the above discussed works (Kamalloo et al., 2023; Liu et al., 2023b), and our method without using the image is similar to what is used in (Kamalloo et al., 2023).

3.1 Validating `GoEval`

To validate `GoEval`, we first generate outputs on our experimental dataset using Gemini-1.5-Pro (Team et al., 2024), a state-of-the-art VLM. Then, we manually evaluate all the answers on the validation set, marking 0 for incorrect answer and 1 for correct answer. These are the results of human evaluation.

We compare BERTScore precision, recall, F1-score, and six variants of `GoEval` using GPT-4o and GPT-4o mini (OpenAI et al., 2024), with and without reference answers, and with and without using images with human evaluations in terms of accuracy and Kendall Tau. The exact prompts that we use are outlined in Table 8 of Appendix A. When we don't use the image, we also alter the template text by a little to not ask model to refer the image. The results are detailed in Table 3.

Method	Acc (%)	τ
BERTScore-p	1.30	36.21
BERTScore-r	1.30	12.68
BERTScore-f1	1.30	28.50
GoEval (-R, -I)	49.91	16.17
GoEval-mini (-R, -I)	52.70	10.84
GoEval (-R)	71.43	35.94
GoEval-mini (-R)	64.01	24.37
GoEval	78.48	52.43
GoEval-mini	**80.33**	**56.71**

Table 3: Comparison of different evaluation methods: Accuracy (Acc.) and correlation (τ for Kendall's Tau) of evaluation metrics with human judgement. -R indicates absence of reference answers, -I indicates absence of image from the request. `GoEval` with all components (reference answers, image) performs the best.

From the results, we can see that `GoEval` with reference answer and image provides the best alignment with human judgment. Without using the image (-R, -I), the performance proves to be very weak, depicting that existing metrics that do not utilize the image will perform poorly on VQA. We believe this is because image context is crucial for VQA, and in the absence of it, the model isn't able to reason well whether the provided answer is correct or not. Moreover, the differences between models are largely influenced by the amount of information. For example, performance of `GoEval` and `GoEval`-mini with image but without reference only has a difference of $\Delta\tau = 11$, but as soon as the reference answers are added, τ increases by 20 units. This is because the reference answers serve as an extra guidance in addition of the image to determine the correctness of the candidate answer, making the model perform better.

In summary, `GoEval` shows high accuracy and Kendall's Tau correlation with human evaluation, and `GoEval`-mini marginally outperforms `GoEval`. This complements `VQA360` to provide an end-to-end cohesive framework for VLM evaluation in entirely multimodal settings.

4 Comparative Evaluation of VLMs

We evaluate state-of-the-art (SOTA) VLMs in practical settings using our framework using `GoEval`-mini, since it demonstrated maximum correlation with human judgments, and is more cost-efficient.

For VLMs, we use InternVL-2 1B and 8B (Chen et al., 2023, 2024), PaliGemma-3B (Beyer* et al., 2024), Qwen-2-VL 2B and 7B (Bai et al., 2023; team, 2024), LlaVa-1.6-Mistral-7B (Liu et al., 2024), CogVLM2-Llama-3-19B (Wang et al., 2023; Hong et al., 2024), Gemini-1.5 Flash and Pro (Team et al., 2024), and GPT-4o-Mini (OpenAI et al., 2024). The rationale behind choosing these models is to have sufficient diversity to allow users to choose the appropriate VLM based on other constraints.

All models except Gemini-1.5 and GPT-4o are open sourced, which gives freedom to customize the models as desired, and host it in-house. It takes away the privacy and regulatory risk of sending data to a third-party, and reduces operational and opportunity cost factor, as these APIs are costly with rate limits. In-house hosting allows a relatively fixed cost. We have taken smaller models in the 1B-3B range which can be used in resource-constrained environments and on-device AI.

For all open-source models, we use the Hugging-Face implementation with the image and question in a prompt recommended by the model card, since we want to evaluate all scenarios uniformly. We use Gemini APIs[3] OpenAI API[4] for Gemini and GPT-4o-mini. The results are discussed in the following subsections. More details of the artifacts used are given in Table 5 of the Appendix A.1.

4.1 Correlation Between VLM Outputs

Our first hypothesis was that all VLMs do not perform similarly with all task instances. They have their own strengths and weaknesses. To establish that, we evaluated the correlation between the out-

---

[3]https://ai.google.dev/gemini-api/docs
[4]https://platform.openai.com/docs/overview

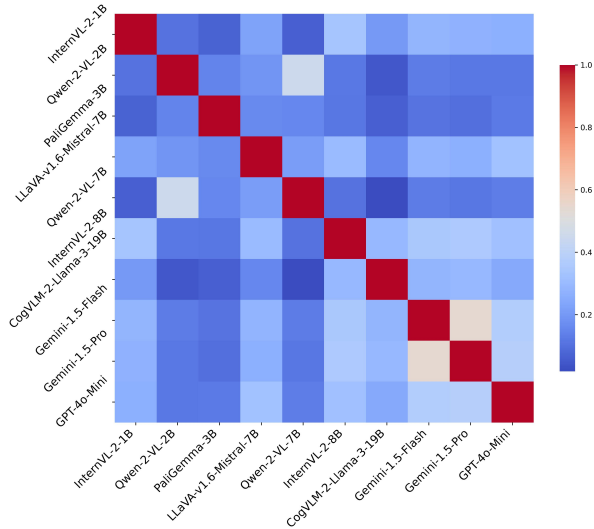puts of different models for all tasks and document the results in Figure 4.

Figure 4: Correlation of `GoEval-mini` values between performance of different VLMs for all task instances. The low correlation values for outputs between all models indicate different VLMs perform differently with task instances.

From the figure, most of the correlations are low. This shows that the performance difference between VLMs is not just in terms of a global statistic, but also differs for individual task instances. Thus, some VLMs that might be great in a subset of tasks may be poor in another subset.

The highest correlation is observed between Gemini-1.5 Flash and Pro, followed by Qwen-2-VL 2B and 7B, and InternVL-2-1B and 8B. Since they are from the same family, they might have similar training data, training strategy, and architecture. Some open-source VLMs also exhibit higher correlations (light blue) with Gemini and GPT, though these correlations remain relatively low.

Since task types, application domains, and knowledge types are key factors in which tasks can be differentiated in practical settings, we move to analyzing the performance on those factors.

4.2 Evaluation on Task Types

We evaluated VLMs on four different task types: chart understanding using ChartQA dataset (Masry et al., 2022), Document Understanding using DocVQA dataset (Mathew et al., 2021), knowledge-based VQA using A-OKVQA (Schwenk et al., 2022) and OK-VQA (Marino et al., 2019), and general VQA using VQAv2 (Goyal et al., 2017a). The performance of VLMs across these is summarized in Table 4.

From the table, the closed models, which are believed to be SOTA, outperform the smaller, open-sourced models. This is expected considering that those models are larger. Within them, we identify that Gemini-1.5-Pro performs significantly better than GPT-4o-mini when extracting information from an image is key, like interpreting documents and charts. On the other hand, where knowledge and comprehension are critical, GPT-4o-mini outperforms Gemini-1.5-Pro. If processing cost is a factor, Gemini-1.5-Flash can be chosen with approximately 7-10% performance loss.

Among open-source models, InternVL-2 shows promising results with both 1B and 8B models given their size, and can be chosen if open-source models are needed. CogVLM-2-Llama-3-19B also competes closely with Gemini-1.5-Flash in Chart and Document understanding tasks. Llava-1.6-Mistral-7B performs acceptable in knowledge-based VQA and VQA, but the performance degrades drastically in the other two categories, where visual comprehension is critical, exposing its limitations in that area. Qwen-2-VL variants and PaliGemma-3B surprisingly prove to be weak in all tasks.

4.3 Evaluation on Application Domains

We evaluated the VLMs in all application domains that had more than 300 task instances, and the results are shown in Figure 5.
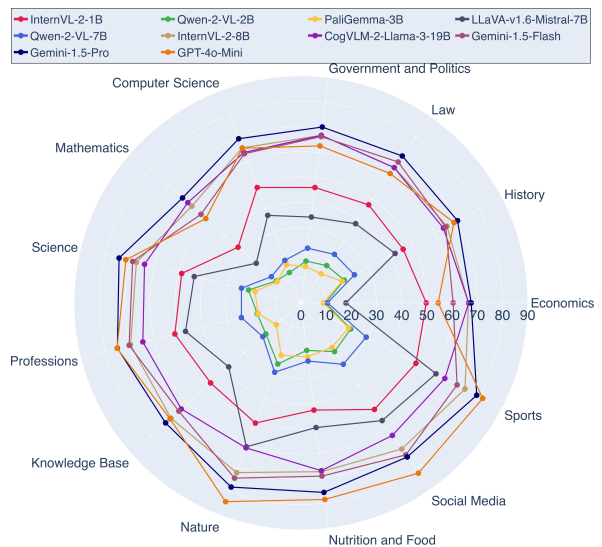

Figure 5: Mean `GoEval-mini` scores for different application domains for all VLMs. Gemini-1.5-Pro and GPT-4o-mini are the best performing closed models, with CogVLM-2-LlaMa-3-19B and InternVL-2-8B performing the best amongst open models.

Task Type	Intern-VL-2-1B	Qwen-2-VL-2B	Pali-Gemma-3B	LLaVA-1.6-Mistral-7B	Qwen-2-VL-7B	Intern-VL-2-8B	Cog-VLM-2-Llama-3-19B	Gemini-1.5-Flash	Gemini-1.5-Pro	GPT-4o-Mini
Chart U.	46.81	12.31	12.84	22.88	14.06	65.68	67.34	62.17	**67.66**	60.36
Document U.	45.68	15.81	12.49	30.07	20.00	65.76	68.17	64.36	**70.53**	54.68
KBQA	45.41	19.83	19.22	58.65	26.94	71.43	62.77	70.66	77.31	**84.04**
VQA	55.81	26.29	23.58	57.64	29.52	68.65	57.18	69.11	73.25	**77.74**

Table 4: Mean `GoEval`-mini scores for different task types for all VLMs. **Bold** numbers indicate best results. Gemini-1.5-Pro performs better in chart and document understanding, while GPT-4o-mini performs better in the other two (U. = Understanding, KBQA = Knowledge-based VQA).

If the VLMs were similar in all application domains, their result would be perfectly circular. However, we see that most of the VLM performance graphs have aberrations in multiple categories, highlighting variance in strengths. In most cases, the weakness of one of the few VLMs is compensated for by the strength of others. So, choosing a VLM wisely according to the application need can help mitigate weaknesses.

Among the closed models, GPT-4o-mini proves to be the best in four categories - Nature, Nutrition and Food, Social Media and Sports, and Gemini-1.5-Pro proves to be the best in all other categories. GPT-4o-mini doesn't even remain second best in some categories like Mathematics, Economics, Law, and is outperformed by many open-sourced models here. Gemini-1.5-Flash remains strong with a performance deficit compared to Pro, but in Social Media tasks, it almost matches Pro. Therefore, while the appropriate model should be selected based on requirements, the Gemini-1.5-Pro generally looks to be the best overall choice.

In the open-source model category, InternVL-2-8B and CogVLM-2-Llama-3-19B are the best possible choices. CogVLM-2-Llama-3-19B generally performs well in more academic topics like Mathematics, Computer Science, law, Government and Politics, but suffers a lot of performance degradation in more social topics like Nature, Nutrition and Food, Social Media, Sports. InternVL-2-8B also shows similar traits, but the difference is relatively less. For academic topics, these models even outperform some of the closed models. Llava-1.6-Mistral-7B is one model that shows exactly opposite trait than this, being limited in academic topics as compared to social topics. Qwen-2 variants and PaliGemma show weak results in all domains, like in task types. InternVL-2-1B remains the best choice if a small model is required, with decent

results using 1B parameters.

4.4 Evaluation on Knowledge Types

Similar to application domains, we evaluate all VLMs on all knowledge types where number of task instances is greater than 300. We demonstrate the results using a similar radar chart in Figure 6.
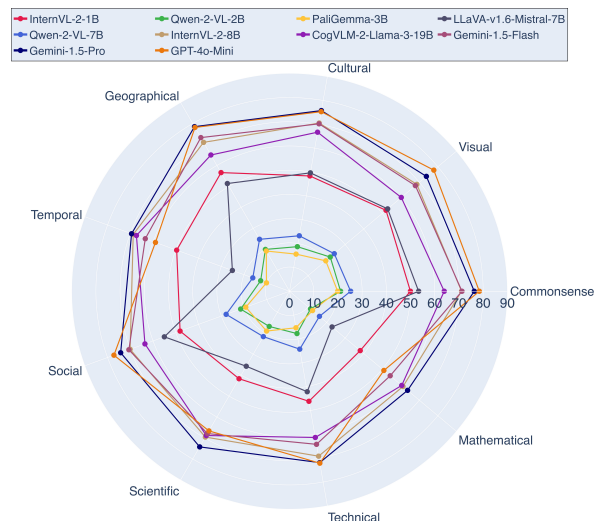

Figure 6: Mean `GoEval`-mini scores for different knowledge types for all VLMs. Gemini-1.5-Pro and GPT-4o-mini perform the best in most knowledge types, while InternVL-2 models also demonstrate competitive performance based on their size.

In the closed models category, the SOTA performance is again shared by Gemini-1.5-Pro and GPT-4o-mini. But unlike in application domains, GPT-4o-mini competes more strongly with Gemini-1.5-Pro on several knowledge types. Visual, Social and Commonsense knowledge is where the advantage of GPT-4o-mini over Gemini-1.5-Pro is maximum, again depicting its strength in social knowledge types. It is still considerably weak in temporal, scientific, and mathematical knowledge, falling behind even some of the small open-source

models. Gemini-1.5-Pro will still be the best overall choice, but different models can be selected based on the results obtained for specific knowledge types. Gemini-1.5-Flash again follows similar trend, with slight performance difference from Pro.

Comparing open-source models, InternVL-2-8B and CogVLM-2-Llama-3-8B are the best choices here as well, but here we can see that InternVL-2-8B outperforms the latter in all knowledge types. InternVL-2-1B continues to be the best overall choice if small models are required, and the Qwen-2 variants and PaliGemma continue to be the least effective. Therefore, InternVL-2-8B is the best overall choice for open source models.

4.5 Overall Analysis and Recommendations

In subsections 4.2, 4.3 and 4.4, we evaluated all vision-language models under three different aspects of task types, application domains and knowledge types. We also discussed their strengths and weaknesses in those categories and recommended different VLMs to use under different requirements. In this section, we will take a higher level look at everything together.

We identified that Gemini-1.5-Pro and GPT-4o-mini are different. In general, GPT-4o-mini is weaker in analytical tasks, like tasks of Mathematics or Economics domain, or scientific or mathematical knowledge types. It not only falls behind Gemini-1.5-Pro in these tasks, but also behind open models like InternVL-2-8B or CogVLM-2-Llama-3-19B. However, it is strong is social and topical tasks like Nature, Sports, Nutrition and Food domains. Gemini-1.5-Pro proves strong more generally, but is expensive. It is either the best, or comes close second or third best. Therefore, if cost is a significant factor, Gemini-1.5-Flash can be considered as a decent alternative at a performance deficit of around 7-10%.

InternVL-2-8B and CogVLM-2-Llama-3-19B are strong open-source models. Due to size differences, resource availability also contributes to deciding which model to use. CogVLM-2-Llama-3-19B is better at more academic tasks, that belong to domains like History, Law, Computer Science, etc., or knowledge types like Temporal, Scientific or Mathematical Knowledge. InternVL-2-8B is a more general capable model that demonstrates more suitability in broader application requirements. In some cases, it outperforms GPT-4o-mini as well. Possessing superior performance in addition to other advantages of open-sourced model

makes it a strong choice. It can also be aligned for downstream tasks to improve performance.

Among small models suited for on-device AI, resource-efficient environments, InternVL-2-1B proves is strongest overall, significantly outperforming models like Qwen2-VL-2B and PaliGemma-3B in all categories.

The Qwen-2-VL variants and PaliGemma-3B did not prove fit for use in our experimental settings, being very weak on all categories. LlaVa-1.6-Mistral-7B also performs average, similar to InternVL-2-1B, but is weak in all aspects compared to InternVL-2-8B, a similar-sized model.

Since Gemini-1.5-Pro was the most successful model, we demonstrate some of the qualitative examples using that model in Table 9 in the Appendix A. Finally, Tables 10 and 11 of the Appendix A contain quantitative results on all application domains and knowledge types, respectively, including the categories that were excluded from the study of the main paper.

5 Conclusion

In this paper, we propose a comprehensive framework for evaluating Vision-Language Models (VLMs) across diverse visual question-answering (VQA) tasks, addressing specific application requirements. Our framework introduces a novel evaluation paradigm that classifies VQA tasks along three dimensions: *task types*, *application domains*, and *knowledge types*. To support this, we release VQA360, a dataset annotated across 4 task types, 22 application domains, and 15 knowledge types, derived from established VQA benchmarks. We also present GoEval, a new evaluation metric to complement it, leveraging GPT-4o to integrate visual and textual information, achieving a 56.71% correlation with human judgments and outperforming traditional metrics.

Through experiments with 10 state-of-the-art VLMs, we observe significant performance variation across categories, with no single model proving universally optimal. Proprietary models like Gemini-1.5-Pro and GPT-4o-mini achieve the highest overall performance, while open-source models such as InternVL-2-8B and CogVLM-2-Llama-3-19B excel in specific scenarios. Our findings provide actionable insights for task-specific VLM selection, and establishes a evaluation framework that can be extended to other vision-language tasks, fostering progress in multimodal research.

References

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 (ICCV)*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Lucas Beyer*, Andreas Steiner*, André Susano Pinto*, Alexander Kolesnikov*, Xiao Wang*, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai*. 2024. PaliGemma: A versatile 3B VLM for transfer. *arXiv preprint arXiv:2407.07726*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3558–3568.

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

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *Preprint*, arXiv:2307.08691.

Xingyu Fu, Sheng Zhang, Gukyeong Kwon, Pramuditha Perera, Henghui Zhu, Yuhao Zhang, Alexander Hanbo Li, William Yang Wang, Zhiguo Wang, Vittorio Castelli, Patrick Ng, Dan Roth, and Bing Xiang. 2023. Generate then select: Open-ended visual question answering guided by world knowledge. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2333–2346, Toronto, Canada. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017a. 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 (CVPR)*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017b. 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 (CVPR)*.

Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. 2024. Cogvlm2: Visual language models for image and video understanding. *Preprint*, arXiv:2408.16500.

Kushal Kafle and Christopher Kanan. 2017. An analysis of visual question answering algorithms. *Preprint*, arXiv:1703.09684.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada. Association for Computational Linguistics.

Shengzhi Li and Nima Tajbakhsh. 2023. Scigraphqa: A large-scale synthetic multi-turn question-answering dataset for scientific graphs. *Preprint*, arXiv:2308.03349.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26689–26699.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

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

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

Manoj Mathew, Shijian Lu, Antonio Torralba, and Dimosthenis Karatzas. 2021. Document visual question answering (docvqa). In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2200–2209.

OpenAI. 2023. Gpt-3.5 turbo. OpenAI API. Available from OpenAI: https://platform.openai.com/docs/models/gpt-3.5-turbo.

OpenAI. 2024. Gpt-4o: Large language model. Accessed: 2024-09-06.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner,

Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. *Preprint*, arXiv:2206.01718.

Neelabh Sinha, Vinija Jain, and Aman Chadha. 2024. Are small language models ready to compete with large language models for practical applications? *Preprint*, arXiv:2406.11402.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya, Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud, Linda Friso, Abhimanyu Goyal, Ben Caine, Kris Cao, Da-Woon Chung, Matthew Lamm, Gabe Barth-Maron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Orgil Batsaikhan, Mateo Wirth, James Qin, Ivo Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurumurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya Attaluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine, Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillicrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette, Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang, Thi Avrahami, Ye Zhang, Emanuel Taropa, Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano, HyunJeong Choe, Alex Tomala, Chalence Safranek-Shrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire Fortunato, Christina Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakićević, Dominik Grewe, Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Sid Dalmia, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong, Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny Huang, Vedant Misra, Jeremy Chen, RJ Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkelsson, Marcello Maggioni, Daniel Zheng, Yury Sulsky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer, Lars Lowe Sjos, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu

86

Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh Newlan, Zeyncep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlas, Ada Ma, Juraj Gottweis, Jinwei Xing, Chenjie Gu, Jin Miao, Christian Frank, Zeynep Cankara, Sanjay Ganapathy, Ishita Dasgupta, Steph Hughes-Fitt, Heng Chen, David Reid, Keran Rong, Hongmin Fan, Joost van Amersfoort, Vincent Zhuang, Aaron Cohen, Shixiang Shane Gu, Anhad Mohananey, Anastasija Ilic, Taylor Tobin, John Wieting, Anna Bortsova, Phoebe Thacker, Emma Wang, Emily Caveness, Justin Chiu, Eren Sezener, Alex Kaskasoli, Steven Baker, Katie Millican, Mohamed Elhawaty, Kostas Aisopos, Carl Lebsack, Nathan Byrd, Hanjun Dai, Wenhao Jia, Matthew Wiethoff, Elnaz Davoodi, Albert Weston, Lakshman Yagati, Arun Ahuja, Isabel Gao, Golan Pundak, Susan Zhang, Michael Azzam, Khe Chai Sim, Sergi Caelles, James Keeling, Abhanshu Sharma, Andy Swing, YaGuang Li, Chenxi Liu, Carrie Grimes Bostock, Yamini Bansal, Zachary Nado, Ankesh Anand, Josh Lipschultz, Abhijit Karmarkar, Lev Proleev, Abe Ittycheriah, Soheil Hassas Yeganeh, George Polovets, Aleksandra Faust, Jiao Sun, Alban Rrustemi, Pen Li, Rakesh Shivanna, Jeremiah Liu, Chris Welty, Federico Lebron, Anirudh Baddepudi, Sebastian Krause, Emilio Parisotto, Radu Soricut, Zheng Xu, Dawn Bloxwich, Melvin Johnson, Behnam Neyshabur, Justin Mao-Jones, Renshen Wang, Vinay Ramasesh, Zaheer Abbas, Arthur Guez, Constant Segal, Duc Dung Nguyen, James Svensson, Le Hou, Sarah York, Kieran Milan, Sophie Bridgers, Wiktor Gworek, Marco Tagliasacchi, James Lee-Thorp, Michael Chang, Alexey Guseynov, Ale Jakse Hartman, Michael Kwong, Ruizhe Zhao, Sheleem Kashem, Elizabeth Cole, Antoine Miech, Richard Tanburn, Mary Phuong, Filip Pavetic, Sebastien Cevey, Ramona Comanescu, Richard Ives, Sherry Yang, Cosmo Du, Bo Li, Zizhao Zhang, Mariko Iinuma, Clara Huiyi Hu, Aurko Roy, Shaan Bijwadia, Zhenkai Zhu, Danilo Martins, Rachel Saputro, Anita Gergely, Steven Zheng, Dawei Jia, Ioannis Antonoglou, Adam Sadovsky, Shane Gu, Yingying Bi, Alek Andreev, Sina Samangooei, Mina Khan, Tomas Kocisky, Angelos Filos, Chintu Kumar, Colton Bishop, Adams Yu, Sarah Hodkinson, Sid Mittal, Premal Shah, Alexandre Moufarek, Yong Cheng, Adam Bloniarz, Jaehoon Lee, Pedram Pejman, Paul Michel, Stephen Spencer, Vladimir Feinberg, Xuehan Xiong, Nikolay Savinov, Charlotte Smith, Siamak Shakeri, Dustin Tran, Mary Chesus, Bernd Bohnet, George Tucker, Tamara von Glehn, Carrie Muir, Yiran Mao, Hideto Kazawa, Ambrose Slone, Kedar Soparkar, Disha Shrivastava, James Cobon-Kerr, Michael Sharman, Jay Pavagadhi, Carlos Araya, Karolis Misiunas, Nimesh Ghelani, Michael Laskin, David Barker, Qiujia Li, Anton Briukhov, Neil Houlsby, Mia Glaese, Balaji Lakshminarayanan, Nathan Schucher, Yunhao Tang, Eli Collins, Hyeontaek Lim, Fangxiaoyu Feng, Adria Recasens, Guangda Lai, Alberto Magni, Nicola De Cao, Aditya Siddhant, Zoe Ashwood, Jordi Orbay, Mostafa Dehghani, Jenny Brennan, Yifan He, Kelvin Xu, Yang Gao, Carl Saroufim, James Molloy, Xinyi Wu, Seb Arnold, Solomon Chang, Julian Schrittwieser, Elena Buchatskaya, Soroush Radpour, Martin Polacek, Skye Giordano, Ankur Bapna, Simon Tokumine, Vincent Hellendoorn, Thibault Sottiaux, Sarah Cogan, Aliaksei Severyn, Mohammad Saleh, Shantanu Thakoor, Laurent Shefey, Siyuan Qiao, Meenu Gaba, Shuo yiin Chang, Craig Swanson, Biao Zhang, Benjamin Lee, Paul Kishan Rubenstein, Gan Song, Tom Kwiatkowski, Anna Koop, Ajay Kannan, David Kao, Parker Schuh, Axel Stjerngren, Golnaz Ghiasi, Gena Gibson, Luke Vilnis, Ye Yuan, Felipe Tiengo Ferreira, Aishwarya Kamath, Ted Klimenko, Ken Franko, Kefan Xiao, Indro Bhattacharya, Miteyan Patel, Rui Wang, Alex Morris, Robin Strudel, Vivek Sharma, Peter Choy, Sayed Hadi Hashemi, Jessica Landon, Mara Finkelstein, Priya Jhakra, Justin Frye, Megan Barnes, Matthew Mauger, Dennis Daun, Khuslen Baatarsukh, Matthew Tung, Wael Farhan, Henryk Michalewski, Fabio Viola, Felix de Chaumont Quitry, Charline Le Lan, Tom Hudson, Qingze Wang, Felix Fischer, Ivy Zheng, Elspeth White, Anca Dragan, Jean baptiste Alayrac, Eric Ni, Alexander Pritzel, Adam Iwanicki, Michael Isard, Anna Bulanova, Lukas Zilka, Ethan Dyer, Devendra Sachan, Srivatsan Srinivasan, Hannah Muckenhirn, Honglong Cai, Amol Mandhane, Mukarram Tariq, Jack W. Rae, Gary Wang, Kareem Ayoub, Nicholas FitzGerald, Yao Zhao, Woohyun Han, Chris Alberti, Dan Garrette, Kashyap Krishnakumar, Mai Gimenez, Anselm Levskaya, Daniel Sohn, Josip Matak, Inaki Iturrate, Michael B. Chang, Jackie Xiang, Yuan Cao, Nishant Ranka, Geoff Brown, Adrian Hutter, Vahab Mirrokni, Nanxin Chen, Kaisheng Yao, Zoltan Egyed, Francois Galilee, Tyler Liechty, Praveen Kallakuri, Evan Palmer, Sanjay Ghemawat, Jasmine Liu, David Tao, Chloe Thornton, Tim Green, Mimi Jasarevic, Sharon Lin, Victor Cotruta, Yi-Xuan Tan, Noah Fiedel, Hongkun Yu, Ed Chi, Alexander Neitz, Jens Heitkaemper, Anu Sinha, Denny Zhou, Yi Sun, Charbel Kaed, Brice Hulse, Swaroop Mishra, Maria Georgaki, Sneha Kudugunta, Clement Farabet, Izhak Shafran, Daniel Vlasic, Anton Tsitsulin, Rajagopal Ananthanarayanan, Alen Carin, Guolong Su, Pei Sun, Shashank V, Gabriel Carvajal, Josef Broder, Iulia Comsa, Alena Repina, William Wong, Warren Weilun Chen, Peter Hawkins, Egor Filonov, Lucia Loher, Christoph Hirnschall, Weiyi Wang, Jingchen Ye, Andrea Burns, Hardie Cate, Diana Gage Wright, Federico Piccinini, Lei Zhang, Chu-Cheng Lin, Ionel Gog, Yana Kulizhskaya, Ashwin Sreevatsa, Shuang Song, Luis C. Cobo, Anand Iyer, Chetan Tekur, Guillermo Garrido, Zhuyun Xiao, Rupert Kemp, Huaixiu Steven Zheng, Hui Li, Ananth Agarwal, Christel Ngani, Kati Goshvadi, Rebeca Santamaria-Fernandez, Wojciech Fica, Xinyun Chen, Chris Gorgolewski, Sean Sun, Roopal Garg, Xinyu Ye, S. M. Ali Eslami, Nan Hua, Jon Simon, Pratik Joshi, Yelin Kim, Ian Tenney, Sahitya Potluri, Lam Nguyen Thiet, Quan Yuan, Florian Luisier, Alexandra Chronopoulou, Salvatore Scellato, Praveen Srinivasan, Minmin Chen, Vinod Koverkathu, Valentin Dalibard, Yaming Xu, Brennan Saeta, Keith Anderson, Thibault Sellam, Nick Fernando, Fantine Huot, Junehyuk Jung, Mani

Varadarajan, Michael Quinn, Amit Raul, Maigo Le, Ruslan Habalov, Jon Clark, Komal Jalan, Kalesha Bullard, Achintya Singhal, Thang Luong, Boyu Wang, Sujeevan Rajayogam, Julian Eisenschlos, Johnson Jia, Daniel Finchelstein, Alex Yakubovich, Daniel Balle, Michael Fink, Sameer Agarwal, Jing Li, Dj Dvijotham, Shalini Pal, Kai Kang, Jaclyn Konzelmann, Jennifer Beattie, Olivier Dousse, Diane Wu, Remi Crocker, Chen Elkind, Siddhartha Reddy Jonnalagadda, Jong Lee, Dan Holtmann-Rice, Krystal Kallarackal, Rosanne Liu, Denis Vnukov, Neera Vats, Luca Invernizzi, Mohsen Jafari, Huanjie Zhou, Lilly Taylor, Jennifer Prendki, Marcus Wu, Tom Eccles, Tianqi Liu, Kavya Kopparapu, Francoise Beaufays, Christof Angermueller, Andreea Marzoca, Shourya Sarcar, Hilal Dib, Jeff Stanway, Frank Perbet, Nejc Trdin, Rachel Sterneck, Andrey Khorlin, Dinghua Li, Xihui Wu, Sonam Goenka, David Madras, Sasha Goldshtein, Willi Gierke, Tong Zhou, Yaxin Liu, Yannie Liang, Anais White, Yunjie Li, Shreya Singh, Sanaz Bahargam, Mark Epstein, Sujoy Basu, Li Lao, Adnan Ozturel, Carl Crous, Alex Zhai, Han Lu, Zora Tung, Neeraj Gaur, Alanna Walton, Lucas Dixon, Ming Zhang, Amir Globerson, Grant Uy, Andrew Bolt, Olivia Wiles, Milad Nasr, Ilia Shumailov, Marco Selvi, Francesco Piccinno, Ricardo Aguilar, Sara McCarthy, Misha Khalman, Mrinal Shukla, Vlado Galic, John Carpenter, Kevin Villela, Haibin Zhang, Harry Richardson, James Martens, Matko Bosnjak, Shreyas Rammohan Belle, Jeff Seibert, Mahmoud Alnahlawi, Brian McWilliams, Sankalp Singh, Annie Louis, Wen Ding, Dan Popovici, Lenin Simicich, Laura Knight, Pulkit Mehta, Nishesh Gupta, Chongyang Shi, Saaber Fatehi, Jovana Mitrovic, Alex Grills, Joseph Pagadora, Dessie Petrova, Danielle Eisenbud, Zhishuai Zhang, Damion Yates, Bhavishya Mittal, Nilesh Tripuraneni, Yannis Assael, Thomas Brovelli, Prateek Jain, Mihajlo Velimirovic, Canfer Akbulut, Jiaqi Mu, Wolfgang Macherey, Ravin Kumar, Jun Xu, Haroon Qureshi, Gheorghe Comanici, Jeremy Wiesner, Zhitao Gong, Anton Ruddock, Matthias Bauer, Nick Felt, Anirudh GP, Anurag Arnab, Dustin Zelle, Jonas Rothfuss, Bill Rosgen, Ashish Shenoy, Bryan Seybold, Xinjian Li, Jayaram Mudigonda, Goker Erdogan, Jiawei Xia, Jiri Simsa, Andrea Michi, Yi Yao, Christopher Yew, Steven Kan, Isaac Caswell, Carey Radebaugh, Andre Elisseeff, Pedro Valenzuela, Kay McKinney, Kim Paterson, Albert Cui, Eri Latorre-Chimoto, Solomon Kim, William Zeng, Ken Durden, Priya Ponnapalli, Tiberiu Sosea, Christopher A. Choquette-Choo, James Manyika, Brona Robenek, Harsha Vashisht, Sebastien Pereira, Hoi Lam, Marko Velic, Denese Owusu-Afriyie, Katherine Lee, Tolga Bolukbasi, Alicia Parrish, Shawn Lu, Jane Park, Balaji Venkatraman, Alice Talbert, Lambert Rosique, Yuchung Cheng, Andrei Sozanschi, Adam Paszke, Praveen Kumar, Jessica Austin, Lu Li, Khalid Salama, Wooyeol Kim, Nandita Dukkipati, Anthony Baryshnikov, Christos Kaplanis, Xiang-Hai Sheng, Yuri Chervonyi, Caglar Unlu, Diego de Las Casas, Harry Askham, Kathryn Tunyasuvunakool, Felix Gimeno, Siim Poder, Chester Kwak, Matt Miecnikowski, Vahab Mirrokni, Alek Dimitriev,

Aaron Parisi, Dangyi Liu, Tomy Tsai, Toby Shevlane, Christina Kouridi, Drew Garmon, Adrian Goedeckemeyer, Adam R. Brown, Anitha Vijayakumar, Ali Elqursh, Sadegh Jazayeri, Jin Huang, Sara Mc Carthy, Jay Hoover, Lucy Kim, Sandeep Kumar, Wei Chen, Courtney Biles, Garrett Bingham, Evan Rosen, Lisa Wang, Qijun Tan, David Engel, Francesco Pongetti, Dario de Cesare, Dongseong Hwang, Lily Yu, Jennifer Pullman, Srini Narayanan, Kyle Levin, Siddharth Gopal, Megan Li, Asaf Aharoni, Trieu Trinh, Jessica Lo, Norman Casagrande, Roopali Vij, Loic Matthey, Bramandia Ramadhana, Austin Matthews, CJ Carey, Matthew Johnson, Kremena Goranova, Rohin Shah, Shereen Ashraf, Kingshuk Dasgupta, Rasmus Larsen, Yicheng Wang, Manish Reddy Vuyyuru, Chong Jiang, Joana Ijazi, Kazuki Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemniy, Kiam Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadsy, Prakash Shroff, Inderjit Dhillon, Tejasi Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jiageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush Dubey, Abhi Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang Nguyen, Ji Liu, Madhavi Sewak, Bryce Petrini, DongHyun Choi, Ivan Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chakladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya Samangouei, Riham Mansour, Tomasz Kępa, François-Xavier Aubet, Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Merey, Martin Baeuml, Trevor Strohman, Junwen Bai, Slav Petrov, Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Qwen team. 2024. Qwen2-vl.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. Yfcc100m: the new data in multimedia research. *Communications of the ACM*, 59(2):64–73.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang,

Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. Cogvlm: Visual expert for pretrained language models. *Preprint*, arXiv:2311.03079.

Zihao Wei, Zixuan Pan, and Andrew Owens. 2024. Efficient vision-language pre-training by cluster masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26815–26825.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

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 (CVPR)*.

A Appendix

In this appendix, we provide additional details and results related to our work. The implementation details can be found in Section A.1. Table 7 provides prompts designed to classify tasks by domain (e.g., "Anthropology," "Computer Science") and by the type of knowledge required (e.g., "Commonsense Knowledge," "Visual Knowledge"). Table 8 presents the prompts used in GoEval to verify whether a candidate answer is correct, with or without reference answers, both for visual and text-only evaluation. Additionally, Table 9 details some qualitative examples using the best overall model that we found - Gemini-1.5-Pro. Finally, Tables 10 and 11 contain quantitative results on all application domains and knowledge types, respectively, including the categories that were excluded from the study of the main paper.

A.1 Implementation Details

In this subsection, we will discuss more details around implementation. Table 5 contains all the model cards, which contain exact details of how we implemented all the VLMs, and the recommended prompt templates that we used in our evaluation.

For all the models, we prepend the question with a static text saying 'Only answer the below question. Do not provide any additional information. In addition, we resize all images to 448×448 before sending them through the model. This is the input to every model. For decoding the output, we use Greedy sampling, since fluency is not a key factor in VQA as long as the answers are correct. We use $max\_new\_tokens = 2048$ for all models.

Artifact	Link
VQAv2	Dataset Card
OK-VQA	Dataset Card
A-OKVQA	Dataset Card
ChartQA	Dataset Card
DocumentVQA	Dataset Card
InternVL-2-1B	Model Card
Qwen-2-VL-2B	Model Card
PaliGemma-3B	Model Card
Qwen-2-VL-7B	Model Card
LLaVA-v1.6-Mistral-7B	Model Card
InternVL-2-8B	Model Card
CogVLM-2-Llama-3-19B	Model Card
Gemini-1.5-Flash	Model Card
Gemini-1.5-Pro	Model Card
GPT-4o-Mini	Model Card
BERTScore	Doc (used using Roberta Large)

Table 5: Details of artifacts used with artifact links.

The implementation of executing all models and our evaluation metric can be found in the Code provided. We also provide the implementation of captions and object tags that can be used if this framework is being adapted to other tasks. All the configuration parameters and hardware used are detailed in Table 6.

Configuration Parameter	Specification
Number of GPUs	1, 2 for CogVLM-2
GPU Model	Nvidia A40
GPU Memory Capacity	48 GB
Batch Size	8
Image Resolution	448 × 448 (224 for PaliGemma)
Maximum New Tokens	2048

Table 6: Hardware and model configuration details used in the experiments, highlighting specialized settings for certain models.

We also use 4-bit quantization and Flash Attention 2 (Dao, 2023) wherever supported for memory and execution efficiency.

A.2 Using this Work to Select VLM

The prerequisite to using this work is to lay down the problem statement and its scope along with other system parameters that should include, but should not be limited to resource availability, data availability, system constraints, resource or data processing budget, acceptable performance bounds, etc.

Start with finding the task type, application domain and reasoning type closest to your requirement from Table 4, 10 and 11. Next, from your design constraints, identify some sets of VLMs acceptable for your solution. For example, if using on-device AI, you might only be able to use either small VLMs, or closed model accessible by APIs, depending on your acceptable performance bounds, and regulatory aspects of being able to share data across, having the inference budget for using APIs, and so on. Refer to Section 4 for a more detailed discussion around which models will fit which needs. Between those, check the performance of the subset of categories and subset of models, and choose the best model.

These models provide a comparison on a uniform foundation so that a comparative analysis can be done. These models can further be customized as desired for best outputs as per other design parameters and needs.

Aspect	Prompt to Extract Tags
Prompt to Extract Application Domains	Following are source application domains: Anthropology, Books, Computer Science, Economics, Fiction, Formal logic, Government and Politics, History, Justice, Knowledge Base, Law, Linguistics, Movies, Mathematics, Nature, News, Nutrition and Food, Professions, Public Places, Reviews, Science, Social Media, Sports. There is an image which can be described as: {caption}. The image has the following objects: {object_tags}. A user is asking the following question on the image: {question}, What type of application domain does this task belong to? Choose one or many alternatives from the above options. Return output as list of strings as JSON Object. Example: {{'application_domain': ['domain_a', 'domain_b']}}
Prompt to Extract Knowledge Type	Following are the names and explanation of types of knowledge: Commonsense Knowledge: Knowledge about the world that humans learn from their everyday experiences (e.g., many donuts being made in a cart implies they are for sale rather than for personal consumption). Visual Knowledge: Knowledge of concepts represented visually (e.g., muted color palettes are associated with the 1950s). Cultural Knowledge: Understanding cultural references, norms, and practices (e.g., knowing that a red envelope is associated with good luck in Chinese culture). Temporal Knowledge: Awareness of historical events, timelines, and changes over time (e.g., recognizing a specific style of clothing as being from the 1980s). Geographical Knowledge: Information about locations, landmarks, and regional characteristics (e.g., identifying a famous monument like the Eiffel Tower in Paris). Social Knowledge: Understanding social interactions, relationships, and behaviors (e.g., recognizing that a handshake is a form of greeting). Scientific Knowledge: Knowledge from various scientific domains like physics, biology, chemistry, astronomy, etc. (e.g., understanding that certain plants are poisonous). Technical Knowledge: Familiarity with technology, machinery, and tools (e.g., identifying parts of a computer or types of construction equipment). Mathematical Knowledge: Basic mathematical concepts and their applications (e.g., understanding geometric shapes or calculating areas). Literary Knowledge: Awareness of literature, authors, and genres (e.g., recognizing characters from classic novels). There is an image which can be described as: {caption}. The image has the following objects: {object_tags}. A user is asking the following question on the image: {question}. What type of knowledge is required to answer the question? Choose one or many alternatives from the above options. Return output as list of strings as JSON Object. Example: {{'knowledge_type': ['knowledge_a', 'knowledge_b']}}

Table 7: Prompts used to generate domain and knowledge type tags using the question, image caption and object tags.

Prompt	Reference	Image
Question: {question} Reference Answers: {reference} Candidate Answer: {candidate} Consider Reference Answers to be multiple answers provided for the given question in context with the above image. If there are multiple answers, they are separated by semi-colon(;). Based on the image, is the candidate answer a correct answer for the given question? Answer only 'yes' if the candidate answer is correct or only 'no' if it is not.	✓	✓
Question: {question} Candidate Answer: {candidate} Based on the image, is the candidate answer a correct answer for the given question? Answer only 'yes' if the candidate answer is correct or only 'no' if it is not.	✗	✓
Question: {question} Reference Answers: {reference} Candidate Answer: {candidate} Consider Reference Answers to be multiple answers provided for the given question in context. If there are multiple answers, they are separated by semi-colon(;). Based on the context, is the candidate answer a correct answer for the given question? Answer only 'yes' if the candidate answer is correct or only 'no' if it is not.	✓	✗
Question: {question} Candidate Answer: {candidate} Based on the context, is the candidate answer a correct answer for the given question? Answer only 'yes' if the candidate answer is correct or only 'no' if it is not.	✗	✗

Table 8: Prompts used with GoEval in different settings. (✓) represents included, (✗) represents excluded.
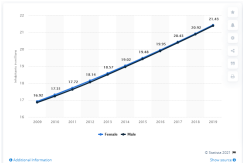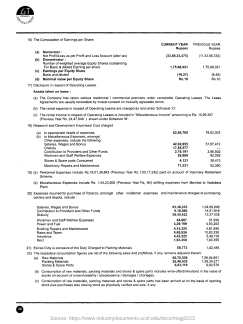
Image	Question	Labels	Response	Human Evaluated Label	GoEval-mini
	The cat is doing what?	laying down, sleeping, laying down, laying, sleeping, resting, laying, laying, laying down, resting	Lying down	✓	✓
	What number is closest to the number of people that are pushing the bus?	eight, eight, eight, ten, twelve, nine, twenty, ten, nine, 15	10	×	✓
	Is there any difference between the male and female values?	No	No	✓	×
	What is the name of the company mentioned in logo?	golden tobacco limited, Golden Tobacco Limited	The logo	×	×
	What is the message written on?	sign, sign, metal sign, stop sign, signs, street sign, sign, metal sign, street sign	A sign	✓	✓

Table 9: Some qualitative examples with Gemini-1.5-Pro. (✓) represents correct answer, saying that response correctly answers the question, (×) represents incorrect answer.

Domain	Intern-VL-2-1B	Qwen-2-VL-2B	Pali-Gemma-3B	LLaVA-v1.6-Mistral-7B	Qwen-2-VL-7B	Intern-VL-2-8B	Cog-VLM-2-Llama-3-19B	Gemini-1.5-Flash	Gemini-1.5-Pro	GPT-4o-Mini
Anthropology	46.47	20.33	16.60	60.58	21.99	73.86	63.18	71.91	74.69	**82.16**
Books	55.56	21.37	10.68	40.60	20.94	70.51	66.23	63.79	**72.22**	66.95
Computer Science	48.96	12.86	16.18	37.14	18.05	65.77	63.83	63.29	**69.65**	65.61
Economics	49.73	9.73	8.92	17.84	10.54	66.49	66.58	60.39	**67.57**	54.47
Fiction	53.70	22.22	24.07	61.11	29.63	81.48	57.41	62.96	70.37	**83.33**
Formal logic	31.96	23.71	14.43	37.50	22.68	56.70	50.00	58.33	**63.92**	58.76
Government and Politics	46.11	16.71	14.57	34.30	21.86	66.96	66.96	66.45	**70.28**	62.72
History	45.80	19.42	18.47	42.21	23.98	65.47	63.86	64.88	**70.19**	68.56
Justice	41.07	23.21	10.71	46.43	32.14	66.07	62.50	64.81	**73.21**	71.43
Knowledge Base	47.91	18.60	13.05	38.30	20.20	68.97	63.44	64.76	**71.78**	69.40
Law	47.29	17.98	14.04	38.18	23.40	65.27	65.17	67.92	**70.79**	62.34
Linguistics	48.00	17.78	20.89	44.89	21.33	67.56	58.04	63.64	75.00	**79.46**
Mathematics	33.33	13.12	12.69	23.71	15.70	57.85	60.00	52.99	**62.80**	50.44
Movies	53.80	22.78	20.89	51.90	32.91	69.62	63.29	66.24	76.58	**78.06**
Nature	50.98	25.92	22.22	61.05	29.41	71.99	61.50	74.34	78.14	**84.32**
News	54.59	17.84	11.35	31.89	17.84	68.11	68.51	70.79	**71.74**	62.78
Nutrition and Food	42.83	18.99	21.55	49.81	23.26	67.44	67.05	69.22	75.73	**78.56**
Other	48.46	19.39	23.64	52.96	24.35	68.32	58.91	67.46	71.63	**78.10**
Professions	51.53	17.88	17.55	47.18	24.35	69.65	64.62	70.24	75.03	**75.15**
Public Places	54.48	18.34	21.72	60.34	24.83	77.59	66.55	75.00	82.35	**82.41**
Reviews	52.86	18.57	23.19	50.00	24.29	64.29	60.00	69.12	70.00	**72.46**
Science	48.75	21.42	18.79	43.63	24.31	67.28	63.94	68.78	**74.34**	71.56
Social Media	51.34	23.47	21.52	56.72	29.58	70.42	63.88	73.27	**74.26**	82.06
Sports	51.55	22.33	21.36	60.53	29.26	73.53	64.50	69.98	78.79	**81.46**

Table 10: Mean `GoEval`-mini for various application domains across multiple VLMs. The best result in each domain is represented in **BOLD**. Note, this also includes the domains that were excluded from main paper's analysis because of having less than 300 task instances.

Knowledge Type	Intern-VL-2-1B	Qwen-2-VL-2B	Pali-Gemma-3B	LLaVA-v1.6-Mistral-7B	Qwen-2-VL-7B	Intern-VL-2-8B	Cog-VLM-2-Llama-3-19B	Gemini-1.5-Flash	Gemini-1.5-Pro	GPT-4o-mini
Commonsense	49.95	21.10	19.88	53.33	25.25	71.06	63.91	71.24	76.32	**78.41**
Cultural	48.36	18.70	15.57	49.65	23.26	70.54	66.70	70.21	**75.77**	75.31
Geographical	56.58	19.95	19.20	51.37	24.77	71.00	64.98	73.28	**78.50**	78.12
Literary	44.25	17.42	13.24	41.46	20.91	68.64	71.58	67.97	**73.78**	66.78
Mathematical	38.13	11.33	12.40	22.93	16.13	61.20	60.51	54.30	**63.67**	50.88
Other	41.82	13.64	20.00	45.45	16.36	62.73	57.27	63.30	67.59	**76.85**
Scientific	41.72	16.78	18.98	35.76	21.63	69.54	68.81	67.87	**74.22**	66.59
Social	48.12	21.49	19.23	55.04	27.87	70.09	63.57	70.63	74.24	**77.20**
Technical	46.15	17.74	15.31	42.13	24.29	69.17	61.32	64.19	71.87	**72.10**
Temporal	49.62	12.69	10.14	25.13	16.15	68.72	67.31	63.42	**69.45**	58.95
Visual	52.03	21.97	19.54	52.92	24.13	68.67	60.27	67.87	73.81	**77.82**

Table 11: Mean `GoEval`-mini for all knowledge types across multiple VLMs. The best result in each knowledge type is represented in **BOLD**. Note, this also includes the knowledge types that were excluded from main paper's analysis because of having less than 300 task instances.