What Are We Measuring When We Evaluate Large Vision-Language Models? An Analysis of Latent Factors and Biases

Anonymous ACL submission

Abstract

Vision-language (VL) models pretrained on colossal image-text datasets have attained broad VL competence, which is difficult to evaluate. A common belief is that a small number of VL skills underlie the variety of VL tests. In this paper, we perform a large-scale transfer learning experiment aimed at discovering latent VL skills from data. We reveal interesting characteristics that have important implications to test suite design. First, generation tasks suffer from a length bias, suggesting benchmarks should balance tasks with different output lengths. Second, we demonstrate that factor analysis successfully identifies reasonable yet surprising VL skill factors, suggesting that VL test suites should consider similar analysis. Finally, we present a new dataset, OLIVE, which simulates user instructions in the wild and presents a unique challenge dissimilar to all datasets we tested. Our findings contribute to the design of balanced and broad-coverage vision-language evaluation methods.

1 Introduction

Benefiting from enormous training data, large model sizes, and pretrained large language models, the current generation of vision-language models (VLMs) (*e.g.,* Dai et al. 2023; Zhu et al. 2023; Liu et al. 2023c; Ye et al. 2023; Li et al. 2023a; Awadalla et al. 2023) demonstrate competence in a wide range of tasks, such as visual question-answering, optical character recognition, spatial relation recognition, and so on. However, their broad competence poses a new challenge to the design of evaluation benchmarks, as most previous work focus on evaluating one or a few capabilities, using data from a single distribution and annotation pipeline. As a result, the test data may not be representative of all possible user inputs, causing potential discrepancy between benchmark scores and actual user experiences.

A currently popular evaluation strategy is to test on an ensemble of tasks and report an average score (Bitton et al., 2023; Xu et al., 2023; Liu et al., 2023d; Yu et al., 2023; Li et al., 2023d; Fu et al., 2023). This type of benchmarks is usually justified with a manual categorization of the test tasks, as a benchmark that covers more categories is often believed to be more comprehensive and more capable of measuring broad competence. For example, TouchStone (Bai et al., 2023b) sort tasks into five skills, ranging from visual recognition to visual storytelling. However, most categorizations are based on human intuition and lack support from empirical evidence.

In this paper, we promote an alternative approach that identifies vision-language (VL) capabilities that underlie various tests directly from data. Inspired by the distributional hypothesis (Firth, 1957), we characterize test tasks using neighborhood structures inferred from transfer learning. That is, transfer learning between datasets that follow similar distributions and require similar VL capabilities will likely yield high performance. By analyzing transfer performance between a large number of source and target tasks, we can observe dataset similarity, infer shared VL capabilities, and gain insights into the VL benchmarks.

Specifically, we finetune four popular VLMs with different strengths, BLIP-2 (Li et al., 2023c), Mini-GPT4 (Zhu et al., 2023), LLaVA (Liu et al., 2023c), and mPLUG-Owl (Ye et al., 2023), on 23 training (source) tasks and evaluate them on 29 test (target) tasks. In total, we obtain a total of 2,784 performance measurements. After that, we examine the patterns and conduct Exploratory Factor Analysis, which discovers six interpretable latent factors underlying the measurements.

The analyses reveal a few surprising findings. First, we find that a surface-form property, the average output length, has surprisingly strong influences on transfer performance. This suggests current evaluation results may be affected by this length bias. Second, factor analysis is capable of discovering unexpected yet reasonable factors that explain model performance. For example, we identify factors that separate reading text off images from multi-hop reasoning. These findings have important implications for the design of unbiased and comprehensive VL benchmarks.

Finally, to simulate real-world user instructions, we present a new vision-language dataset, Open-world Language Instruction for Visual-language Evaluation (OLIVE). OLIVE consists of 9,450 images, 30,120 unique instructions and 47,250 responses. Empirically, we show that OLIVE have a transfer profile distinct from all other dataset that we tested and hence provides a test complementary to existing tasks.

We summarize our contributions as follows:

- We promote the approach of discovering VL skills from data and demonstrate factor analysis as a robust and effective technique for this purpose. Our large-scale experiments lead to findings that can inform future design of VLM test suites.
- We introduce a new benchmark, OLIVE, to evaluate open-ended model responses to diverse instructions.

2 Analysis Techniques

The transfer performance from N source (training) tasks to K target (test) tasks on model m is stored as a matrix $A^{(m)} \in \mathbb{R}^{N\times K}$. The performance numbers of different tasks cannot be compared directly due to difference in scales of the evaluation metrics. Therefore, we first normalize the data so that different test performances can be aggregated. Subsequently, we apply Singular Value Decomposition and Factor Analysis. Both techniques may be understood as decomposition of the matrices $A^{(m)}$, albeit with different mathematical formulation.

2.1 Normalization

We obtain the raw performance number $b_{i,j}^{(m)}$ when we train model m on the ith source task and test on the jth target task. We obtain the normalized performance $a_{i,j}^{(m)}$ as

$$a_{i,j}^{(m)} = (b_{i,j}^{(m)} - b_{0,j}^{(m)})/(\max_{j'} b_{i,j'}^{(m)} - b_{0,j}^{(m)}), \quad (1)$$

where $b_{0,j}^{(m)}$ denotes the performance of the pre-trained model m on target task j without finetuning on any source task, which we refer to as the zero-shot performance. If finetuning on source i improves over the zero-shot performance, $a_{i,j}^{(m)}$ is a positive number. Conversely, we have negative transfer from source i, and $a_{i,j}^{(m)}$ is negative. The best source task, which in most cases is the in-domain i.i.d. training task, has $a_{i,j}^{(m)} = 1$. Hence, this normalization separates positive and negative transfers and shows how source tasks perform relative to the in-domain training data. The matrix $A^{(m)}$ has $a_{i,j}^{(m)}$ as its components. After separate normalization, we concatenate the four matrices $A^{(m)}$, corresponding to the four models we finetune, along the source-task dimension (the rows) and obtain the aggregate performance matrix $A \in \mathbb{R}^{4N\times K}$.

2.2 Singular Value Decomposition

Singular Value Decomposition (SVD) is a classic technique for learning distributed representations. Levy and Goldberg (2014) show that SVD produces word embeddings comparable to word2vec (Mikolov et al., 2013). The SVD of matrix A can be written as

$$A = U\Sigma V^{\top} \quad (2)$$

We perform truncated SVD using the D largest singular values. After that, we use $V\Sigma^{1/2} \in \mathbb{R}^{K\times D}$ as the features of the target tasks.

2.3 Factor Analysis

It is a widely held belief that a small number of factors, known as cognitive abilities, underlie human performance on numerous mental activities (Horn and McArdle, 2007). To uncover these latent factors, Spearman (1904) developed the statistical technique of Factor Analysis. For modern treatments, we refer readers to Gorsuch (2014) and Barber (2012).

In this paper, we start with a premise similar to Spearman's, that a small number of VL capabilities are responsible for VLM performance on various test tasks. This belief is in fact implicitly shared, though often not explicitly stated, by most recent VLM evaluation papers (*e.g.,* Bitton et al. 2023; Xu et al. 2023; Liu et al. 2023d; Bai et al. 2023b) that attempt to categorize VLM test scenarios based on intuitive justifications. In contrast, we apply Exploratory Factor Analysis (EFA) to uncover these factors from empirical data.

Mathematically, we treat the ith column of A, $\boldsymbol{a}_i \in \mathbb{R}^{4N}$, as the characteristic of the target task i, which we try to explain with L latent factors,

$$\boldsymbol{a}_i = W\boldsymbol{h}_i + \boldsymbol{\mu} + \boldsymbol{\epsilon}, \tag{3}$$

where $W \in \mathbb{R}^{4N \times L}$ reflects how source tasks load to the L latent factors and $\boldsymbol{h}_i \in \mathbb{R}^L$ reflects how target task i decompose to the latent factors. $\boldsymbol{\mu}$ is the average vector across target tasks and $\boldsymbol{\epsilon}$ is Gaussian noise. EFA differs from PCA by assuming the covariance of $\boldsymbol{\epsilon}$ is diagonal rather than spherical. Note that the above formulation is invariant up to a rotational matrix R, as $W\boldsymbol{h}_i = (WR)(R^\top \boldsymbol{h}_i)$. We apply the Varimax rotation (Kaiser, 1958) to find R so that $\boldsymbol{h}_i$ is as concentrated on a few factors as possible.

Our preliminary analysis suggests that the captioning and VQA tasks are highly correlated and predominantly load onto a single factor, likely indicative of a general VL capability. To isolate and examine other factors, we apply linear regression to control for the influence of the dominant factor. Specifically, we first perform EFA with one factor, so that W becomes a $4N$-by-1 vector $\boldsymbol{w}$. We then perform regression from $\boldsymbol{w}$ to A by solving the following problem,

$$\text{minimize} \quad \left\| A - \boldsymbol{w}\boldsymbol{\beta}^\top - \boldsymbol{\gamma}\mathbf{1}^\top \right\|_F^2, \tag{4}$$

where $\boldsymbol{\beta}, \boldsymbol{\gamma} \in \mathbb{R}^{4N}$ are trainable parameters. After that, we conduct EFA on the residuals, $\bar{A} = A - \boldsymbol{w}\boldsymbol{\beta}^\top - \boldsymbol{\gamma}\mathbf{1}^\top$, which contain information about other factors indicative of more specific VL capabilities than the first factor.

We employ both parallel analysis and Velicer's Minimum Average Partial (MAP) test to determine the optimal number of factors to extract. Parallel analysis compares eigenvalues from our sample correlation matrix against those from random data of the same size, identifying factors that explain more variance than expected by chance. Conversely, Velicer's MAP test evaluates the average squared partial correlation for each possible number of factors, pinpointing where additional factors no longer meaningfully increase variance explanation. Both these methods converge on the decision to extract 6 factors.

3 Source and Target Tasks

We gather 27 publicly available VL datasets and create variations, yielding 23 source tasks and 29 target tasks. We show the full list of tasks in Tab. 1 and describe them below. The performance metrics used are AUC for Hateful Memes, CIDEr (Vedantam et al., 2015) for OpenCQA, OLIVE, and all captioning datasets, and accuracy for the remaining tasks. To focus on end-to-end performance, we do not perform any separate optical character recognition.

Intuitive Category	Task	Source	Target
Image Captioning	COCO Caption	✓	✓
	Flickr30k	✓	✓
	Web CapFilt	✓	✗
	TextCaps	✓	✓
Generic VQA	VQAv2	G	G, MC
Knowledge-based VQA	OK-VQA	G	G, MC
	A-OKVQA	G, MC	G, MC
	ScienceQA	MC	MC
OCR VQA	TextVQA	G	G, MC
	OCR-VQA	G	G, MC
Visual Reasoning	GQA	G	G, MC
	VSR	MC	MC
	IconQA	MC	MC
	CLEVR	✗	G, MC
	RAVEN-FAIR	✗	MC
Classification	Hateful Memes	MC	MC
Humor & Sarcasm	New Yorker Ranking	✗	✓
	New Yorker Explanation	✗	✓
	MORE	✗	✓
Chart Reading	OpenCQA	G	G
	ChartQA	✗	G, MC
Open-ended Generation	OLIVE (Ours)	✓	✓
	LLaVA Conversation	✓	✗
	LLaVA Reasoning	✓	✗
	LLaVA Description	✓	✗
Question Generation (QG)	VQAv2 QG	✓	✗
	OK-VQA QG	✓	✗
	A-OKVQA QG	✓	✗

Table 1: The list of source and target tasks used in experiments. G and MC indicate the generative and multiple-choice versions of the VQA tasks respectively.

Image Captioning. Image captioning is one of the most popular image-text tasks and is commonly used as a pretraining task for VLMs (Chen et al., 2022; Tiong et al., 2022). Here we select two classic datasets: COCO Caption (Lin et al., 2014) and Flickr30k (Young et al., 2014). In addition, we include TextCaps (Sidorov et al., 2020), which involves description of textual content in images. We also include as a source task Web CapFilt, a set of synthetic image captions on a large variety of web images. Web CapFilt was generated by BLIP (Li et al., 2022b) for self-training. We hypothesize that its diversity could be beneficial in transfer learning.

Visual Question-answering (VQA). VQA is another very popular image-text task due to the versatility of the question-answering format. VQAv2 (Goyal et al., 2017) is probably the most prominent VQA benchmark, with more than 200,000 COCO images and 1 million questions. Other variations include knowledge-grounded VQA, OCR VQA, Chart VQA, and so forth, which we discuss below.

Performance measurement in VQA can be tricky, as there are often many correct answers to the same question. As a remedy, we create two target tasks for every VQA dataset. The first is the generative (G) version, which considers an answer to be correct only when it matches exactly one of the ground-truth answers. The second is the multiple-choice (MC) version, where the model chooses one from five options. To convert a generative VQA dataset to the MC version, we create five options for every question. We include at most two correct answers to account for their linguistic variations. After that, we add incorrect choices by sampling answers from other questions and picking those with top-k probabilities according to InstructBLIP (Dai et al., 2023). During inference, we feed all options to the model and choose the option with the highest average word probability as the model prediction.

Knowledge-grounded VQA. These tasks require the model to apply world knowledge not present in the input to answer questions. ScienceQA (Lu et al., 2022) focus on contents of science textbooks. OK-VQA (Marino et al., 2019) is mainly about visual recognition and knowledge recall, whereas A-OKVQA (Schwenk et al., 2022) often needs one additional step of reasoning.

OCR VQA. TextVQA (Singh et al., 2019) and OCR-VQA (Mishra et al., 2019) are two VQA datasets that requires recognition of text on images. OCR-VQA is about reading text from book covers, whereas TextVQA often requires locating an object before reading the text on it.

Chart Reading. OpenCQA (Kantharaj et al., 2022) and ChartQA (Masry et al., 2022) contain questions regarding the content of diagrams and charts. OpenCQA expects descriptive long-form answers, whereas ChartQA is mainly about data extraction and comparison using short answers.

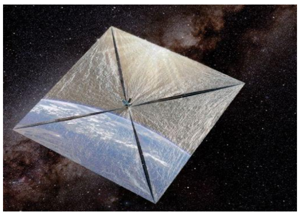
Visual Reasoning. The word reasoning is often used very broadly in the VLM literature. It sometimes refers to shallow tasks like counting (*e.g., how many apples are in the image?*) as well as spatial relations and grounding (*e.g., what is adjacent to the cylinder?*), but could also include logical or algebraic operations. In this category, we include five datasets, GQA (Hudson and Manning, 2019), VSR (Liu et al., 2023a), CLEVR (Johnson et al., 2017), IconQA (Lu et al., 2021), RAVEN-FAIR (Benny et al., 2021). GQA and VSR mainly contain natural images, whereas IconQA contains cartoons and Raven-fair features abstract diagrams. CLEVR contains synthetically rendered images of 3D objects. Among these, RAVEN-FAIR is the most difficult, as the data come from the Raven's test (Raven, 1938), an intelligence test designed for humans that requires complex reasoning.

Image Classification. Hateful Memes (Kiela et al., 2020) is a binary classification task that distinguishes hateful memes from other meme images.

Humor and Sarcasm Understanding. Hessel et al. (2023) show that VLMs perform poorly at humor understanding. From their apper, we adopt New Yorker Ranking, which involves selecting of the best humorous caption for a cartoon from the New Yorker magazine, and New Yorker Explanation, which asks the model to explain why the cartoon and its caption invoke humor. The MORE dataset (Desai et al., 2022) involves explaining why a textual statement associated with a natural image is sarcastic.

Question and Open-ended Generation. We adapt three datasets, VQAv2, OK-VQA, A-OKVQA, for the task of question generation from an image and an answer. Further, we take the three subsets of LLaVA-Instruct-150K (Liu et al., 2023c), which respectively focus on free-form conversation, detailed description and reasoning. We use these as source tasks but not target tasks, since these artificial data may not be representative of real use cases.

OLIVE. Additionally, we include a new dataset, OLIVE, a highly diverse, human-corrected multimodal dataset, which we create to simulate in-the-wild user queries to VLMs. Once a VLM is publicly released, it tends to receive substantially more diverse and idiosyncratic inputs than the available academic datasets. OLIVE is our attempt to



Instruction: What is the item in the image?

Output: The item in the image is a solar sail, which is a device that is designed to harness the energy from sunlight to propel a spacecraft through space without the use of fuel. It is a square shaped piece of cloth that acts like a sail and captures the radiation pressure from the sun to propel the spacecraft forward.

Instruction: What could be the possible cultural significance of this typical tree to the Japanese people?

Output: The cherry blossom tree, also known as "sakura", holds great cultural significance in Japanese society. Its delicate and ephemeral beauty is associated with the transience of life and the seasonal cycle of rebirth and renewal. Every year, Japan celebrates Hanami, a national festival where people gather under the cherry blossom trees to appreciate their beauty and contemplate the impermanence of life.

Instruction: What kind of material is used to create a unique mode of transportation and what other materials could be used?

Output: The motorcycle used wood for the body parts which is a unique material for this transportation. While wood is certainly a unique material for a motorcycle, there are many other materials that could also create a one-of-a-kind mode of transportation. For example, a motorcycle made from recycled materials like metal scraps, old car parts, or discarded plastic could be both sustainable and visually striking.

Figure 1: Examples of the OLIVE benchmark for different categories. From left to right: visual recognition, knowledge-based, and creative writing.

simulate such user queries, so that we may train and evaluate the VLMs under conditions similar to the ultimate use case.

We briefly describe the data curation process here and refer readers to the Appendices A and B for more details. First, we randomly sample 9,450 images from LAION-Aesthetics (Schuhmann et al., 2022), consisting of diverse web images. We take the original LAION caption from the dataset and a few generated image captions from BLIP-2 as the complete image description. Next, we feed customized prompts to ChatGPT (OpenAI, 2023a) to generate instructions conditioned on the image description and five responses to each instruction. After that, a team of data annotators are recruited to verify the data and remove errors, shortcut biases, and harmful content. The procedure yields 30,120 unique instructions and 47,250 responses. We use 6,750 instruction-response pairs for training and another 6,750 for validation, and leave the rest as the test set. Fig. 1 shows some examples[1].

4 Experiments

4.1 Setup

Considering that different VLMs may exhibit different training behaviors, we base our analysis of test data using a four popular VLMs that, for the most part, have not been exposed to the datasets in focus: BLIP-2 (Li et al., 2023c), MiniGPT-4 (Zhu et al., 2023), LLaVA (Liu et al., 2023c), and mPLUG-Owl (Ye et al., 2023). As minor exceptions, BLIP-2 and MiniGPT-4 are pretrained on COCO Caption and Web CapFilt. mPLUG-Owl is exposed to COCO Caption. LLaVA is pretrained on the three LLaVA datasets. We avoid models that have been finetuned on many VQA datasets such as InstructBLIP (Dai et al., 2023), LLaVA 1.5 (Liu et al., 2023b), and Qwen-VL (Bai et al., 2023a).

For each model, we finetune the parameters that are trainable during their respective vision-language pretraining. On each source task, we train for 10K steps with a batch size of 192 for BLIP-2 and 128 for MiniGPT-4, mPLUG-Owl and LLaVA. Other model details and hyperparameters are in the Appendices C and D.

4.2 Results

We defer the transfer performance tables to Appendix E due to space constraints. With these results, we first examine the transfer learning power of source tasks. For every target task, we rank the source tasks in descending order of transfer performance. After that, we compute the harmonic mean of all rankings and show the results in Tab. 3. A-OKVQA (MC), VQAv2, ScienceQA, A-OKVQA and OCR-VQA hold the top-5 positions.

In addition, we examine the effects of output lengths. We partition the source and target tasks into three mutually exclusive and collectively exhaustive groups according to their average output

[1]OLIVE will be released upon acceptance.

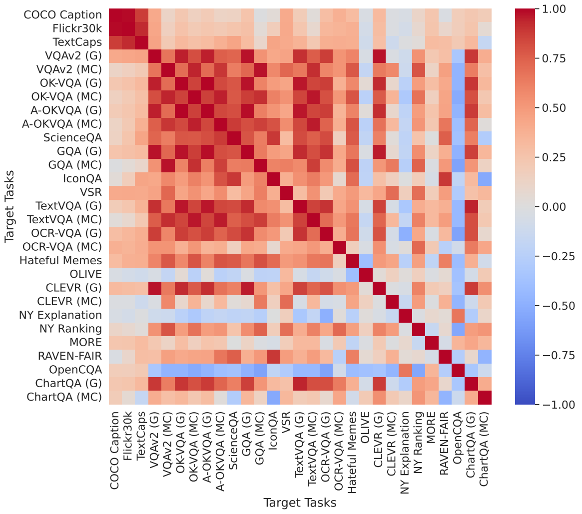


Figure 2: Cosine similarity between target tasks computed using SVD features.

lengths: 1-3 words, 6-12 words, and more than 40 words. We show the average normalized transfer performance across these groups and the top-5 best source tasks for each group in Tab. 2.

Next, we investigate the similarity of target tasks. We perform truncated SVD on A with the first $D = 8$ singular values. After that, we compute cosine similarity between target tasks and visualize the results in Fig. 2. With a mean similarity of -0.06, OLIVE is the third least similar to other target tasks (see details in Appendix F).

Finally, we run EFA on the residual matrix, $\bar{A}$, and present the outcomes in Fig. 3. We plot the most significant factor loading for each target task, retaining only those that exceed an absolute value threshold of 0.3. Loadings close to 1 or -1 signify strong influences of a factor on a target task. Three tasks, New Yorker Explanation and Ranking, and Hateful Memes, do not have loadings more than 0.3 on any discovered factor, suggesting they do not share with the other tasks VL capabilities that can be discovered by EFA. The full results are available in Appendix H.

4.3 Discussion

In this section, we highlight important findings from our experiments.

The output length bias. Tab. 2 demonstrates that the transfer performance is strongly influenced by the output lengths. In the top section, mismatch between the output lengths results in significant performance drops. In the bottom section, the best source tasks almost always have similar output lengths to the target tasks. This surprising finding

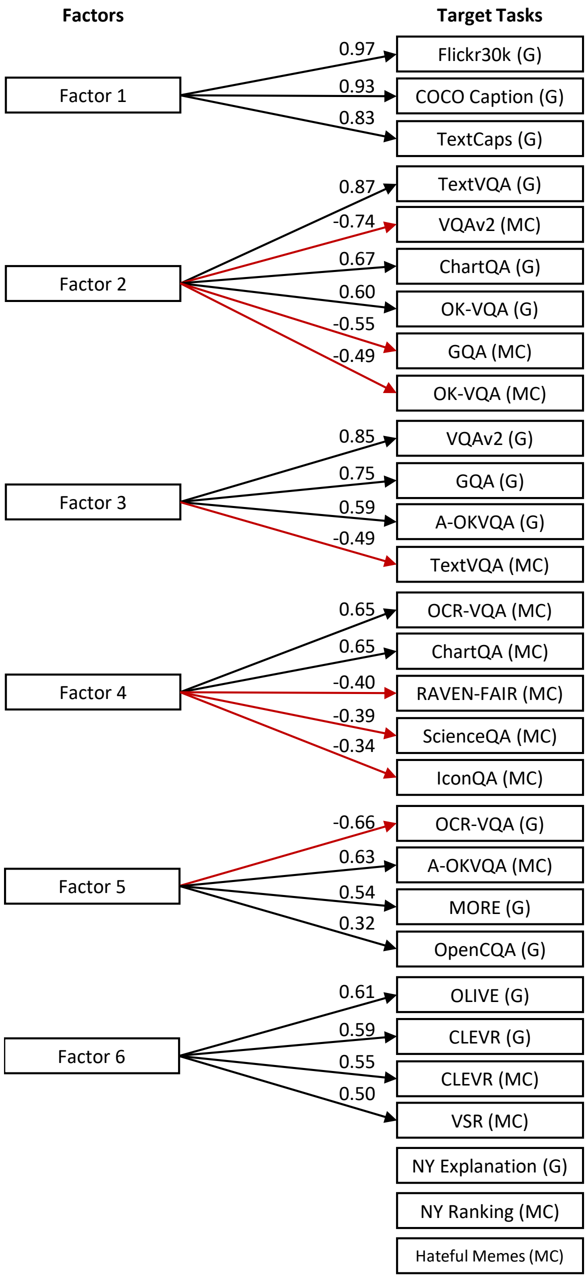


Figure 3: Results of EFA on the residuals $\bar{A}$. Black arrows indicate positive loadings; red arrows indicate negative loadings. Cut-off for factor loadings = 0.3.

shows that output length may be a shortcut feature for VLMs, suggesting that future test suites need a balance among tasks with different output lengths.

EFA Overview. The six EFA factors bear resemblance to hierarchical clusters from SVD features (Appendix G). For example, both techniques identify a factor (Factor 1) or a cluster around the three captioning tasks, COCO, Flickr30k, and TextCaps. However, EFA reveals both the positive and the negative ends of the same factor, which render the factors more interpretable. EFA also picks up more VL skills than SVD, like text reading vs. reasoning.

Source Task Output Length	Target Task Output Length		
	1-3	6-12	>40
1-3	**-0.03 / 1.00**	-0.78 / 0.79	-0.85 / 0.44
6-12	-0.49 / 0.64	-0.43 / **0.75**	**-0.43** / 0.48
>40	-0.90 / 0.43	-0.87 / 0.28	**-0.26 / 0.55**
Top-5 Source Tasks (Mean Length)	VQAv2 (1) A-OKVQA (MC) (1) A-OKVQA (1) OK-VQA (1) TextVQA (1)	Web CapFilt (12) COCO Caption (10) OCR-VQA (3) Flickr30K (12) ScienceQA (3)	LLaVA Conversation (48) OpenCQA (56) TextCaps (12) Flickr30K (12) A-OKVQA (MC) (1)

Table 2: Mean normalized model performance for source and target tasks with different mean output length. In-domain sources are excluded. We reported the mean normalized model performance across all (left) and top-5 (right) source tasks in the output length category.

Source Task	Harmonic Mean Score
A-OKVQA (MC)	1.3
VQAv2 (G)	1.3
ScienceQA (MC)	3.8
A-OKVQA (G)	4.6
OCR-VQA (G)	6.0
GQA (G)	6.2
Flickr30k (G)	7.2
OK-VQA (G)	7.8
Web CapFilt (G)	7.9
IconQA (MC)	8.4
OpenCQA (G)	9.5
TextVQA (G)	9.5
VSR (MC)	10.0
Hateful Memes (MC)	11.7
COCO Caption (G)	13.3
TextCaps (G)	15.0
VQAv2 QG (G)	15.9
LLaVA Conversation (G)	17.6
OLIVE (G)	17.8
A-OKVQA QG (G)	17.9
OK-VQA QG (G)	19.7
LLaVA Reasoning (G)	21.1
LLaVA Description (G)	22.5

Table 3: Harmonic mean ranking score for source tasks.

We note that EFA is affected by the length bias. For example, it does not find a humor factor shared by the two New Yorker tasks. This is because the similarity between the two tasks is masked by their differences in output lengths and MC-vs-G evaluation, resulting in drastically different transfer profiles in Fig. 2. Nonetheless, EFA does find reasonable factors, and we discuss Factors 2-6 below.

Factors 2 and 3: Generative vs. MC evaluation. Factors 2 and 3 capture the contrast between generative and MC evaluation of VQA. Generative VQAs exhibit positive loadings on both factors, whereas MC VQAs load negatively on these factors. Further, generic VQAs, such as VQAv2 and GQA, show negative loadings on Factor 2 and positive loadings on Factor 3. In contrast, VQAs that require specialized OCR capabilities, such as

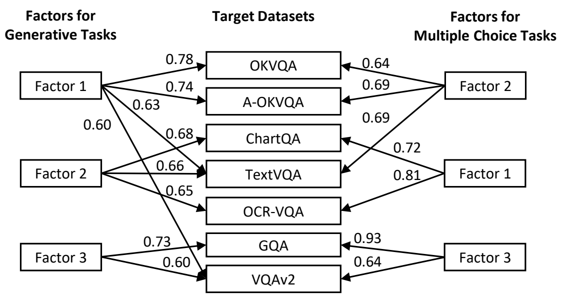


Figure 4: EFA results when we extract 3 factors from the 7 generative VQA tasks and the 7 MC VQA tasks separately. We merge the results for display. Cut-off for factor loadings = 0.6.

TextVQA and ChartQA, load positively on Factor 2 and negatively on Factor 3.

We identify two reasons for the differences between generative evaluation and MC evaluation. First, generative evaluation requires an exact match with at least one ground-truth answer, leading to false negatives on good answers phrased differently. In comparison, the MC evaluation, which compares average word probabilities, does not require strictly word matching. Second, the exact match requirement means that the generative evaluation is more sensitive to the output lengths of the source tasks, which has severe effects as discussed earlier.

Nevertheless, when we analyze factors internal to the generative and MC tasks (Fig. 4), we observe very similar structures. In both groups, we observe a knowledge-based QA factor, which includes OKVQA and A-OKVQA, a OCR-related factor, which includes OCR-VQA and ChartQA, as well as a generic or spatial relation factor, which includes GQA and VQAv2. We observe that EFA can identify robust structures when the input data are conducive.

Factors 4 & 5: Text Reading vs Reasoning. Factors 4 and 5 distinguish between tasks that involve merely text extraction and those that require deeper and multi-hop reasoning. RAVEN-FAIR, ScienceQA, and IconQA, which require strong logical reasoning skills, show negative loadings on Factor 4. Positively associated with Factor 5, A-OKVQA, MORE, and OpenCQA demand reasoning about external knowledge or contextual understanding. In comparison, OCR-VQA and ChartQA mostly involve locating and directly reading text or numbers off the images. The fact that EFA can find these reasonable skills demonstrates the its power.

Factor 6: Spatial reasoning. Factor 6 is characterized by spatial reasoning, as CLEVR and VSR are both designed for this purpose. Notably, while OLIVE shows the highest loading on Factor 6, its communality (overall variance explained) is only 0.4. The remaining variance in OLIVE is not explained by the factors identified in our analysis. This implies that although OLIVE requires spatial reasoning skills, these skills only account for a small portion of skills required by OLIVE.

OLIVE presents a unique test. Intended to simulate real-world user instructions, OLIVE features a unique transfer profile and has very low cosine similarity with other tests (Fig. 2). In addition, EFA only explains 0.4 of the total variance of OLIVE, indicating that the identified factors cannot well explain model behaviors on OLIVE. These facts corroborates our thesis that OLIVE measures an aspect of VL capabilities that few existing datasets test for.

A-OKVQA and VQAv2 are effective source tasks. These two are among the sources with the highest transfer performance (Fig. 3); they transfer well to VQA tasks but also to other complex tasks such as NY Ranking. We hypothesize that the large and diverse data in VQAv2 contributes to its strong transferability. Interestingly, even though A-OKVQA is 24 times smaller than VQAv2, it still transfers well. We hypothesize that the main skill that A-OKVQA is designed for, knowledge-enabled reasoning, is an important skill for VL compentence. In comparison, OK-VQA is designed for only knowledge recall, which is not as beneficial to target tasks.

Humor, sarcasm, and abstract reasoning remain difficult. All models we tested struggle to understand humor and sarcasm, as captured by the New Yorker datasets and MORE. The models also perform barely above chance level on RAVEN-FAIR, an abstract reasoning task. Surprisingly, EFA is able to correctly place RAVEN-FAIR in the reasoning factor (negative Factor 4) despite the tiny variance caused by overall poor performance.

Implications. Our findings have the following implications for the design of VL benchmarks. First, to prevent shortcut learning and giving unfair advantages to any source training task, VL benchmarks should contain tasks of different output lengths and use both generative and MC evaluation. Second, instead of intuition-based categories, VL benchmarks may group tasks based on statistically discovered VL factors, and score VLMs accordingly.

5 Related Work

Relationships among Tasks. Relationships between machine learning tasks has been studied from the perspective of transfer learning and multi-task learning. A number of works focus on identifying transfer relationships from empirical data (Zamir et al., 2018; Achille et al., 2021; Dwivedi and Roig, 2019; Achille et al., 2019; Xi et al., 2023). A typical strategy is to train a base network on multiple source tasks and test the resulting networks on target tasks. In multi-task learning, the research focuses on identifying optimal grouping of tasks that should be learned together to maximize synergy (Standley et al., 2020; Fifty et al., 2021; Ben-David and Schuller, 2003; Kumar and Daume III, 2012; Song et al., 2022). Different from the above, our work focuses on vision-language tasks and identifying latent factors and potential biases responsible for the observed performance.

Broad-coverage Multimodal Test Suites. As VLMs begin to excel on an ever growing set of tasks, the test suites have also grown in size. Early benchmarks contain only a few tasks. For example, Zhou et al. (2022) use 4 tasks and Su et al. (2021) use 8 tasks in 4 groups. More recent benchmarks (Bugliarello et al., 2022; Bitton et al., 2023; Bai et al., 2023b; Yu et al., 2023; Xu et al., 2023; Li et al., 2023d) utilize increasingly more tasks in order to achieve broader coverage of VL capabilities, and group tasks into VL capabilities based on human intuition. For instance, Li et al. (2023b) categorized tasks into 12 aspects focusing on spatial and temporal understanding. To our knowledge, our work is the first data-driven approach to identify the VL capabilities.

6 Conclusions

In this work, we aim to empirically discover latent factors and biases that contribute to the performance of diverse VLMs on VL tasks. Using Exploratory Factor Analysis, we identify six highly interpretable factors, as well as biases that affect performance. We also contribute a new dataset, OLIVE. We hope this research will lead to the creation of VL benchmarks with more balanced and complete coverage of VL capabilities.

7 Ethical Considerations

The potential risks of large language models have been discussed in literature, including Chung et al. (2022), Touvron et al. (2023), and (Chiang et al., 2023). We neither propose nor release new models in this work. For easy replication, we use open-source visions-language models to analyze publicly available, academic datasets.

We consider issues of privacy, toxicity, and fair compensation in the production of OLIVE. We utilize images from LAION-aesthetics, whose privacy policy is in compliance with GDPR. For a given image, we generate an instruction and outputs using ChatGPT. To mitigate the potential issues of hallucination, toxicity, and harmful content in ChatGPT generated content, we hire an annotation company, Flitto, to thoroughly review and correct any errors presented in the data. We establish specific guidelines for annotations, prioritizing content that is free from harmful information. The identity of the annotators are anonymized to safeguard their privacy. We pay Flitto $3 USD per image to ensure fair compensation.

8 Limitations

We focus our study of transfer performance using only one source task instead of multi-task learning setting due to the computational constraint. Thus, we do not investigate the interaction of multiple source tasks on the target tasks. To assess the model transfer performance, our work requires a substantial amount of computation and scaling our approach to diverse models and datasets is inefficient.

References

Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C Fowlkes, Stefano Soatto, and Pietro Perona. 2019. Task2vec: Task embedding for meta-learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6430–6439.

Alessandro Achille, Giovanni Paolini, Glen Mbeng, and Stefano Soatto. 2021. The information complexity of learning tasks, their structure and their distance. *Information and Inference: A Journal of the IMA*, 10(1):51–72.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

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

Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xingxuan Zhang, Junyang Lin, Xinggang Wang, Chang Zhou, and Jingren Zhou. 2023b. Touchstone: Evaluating vision-language models by language models. *arXiv 2308.16890*.

David Barber. 2012. *Bayesian Reasoning and Machine Learning*. Cambridge University Press.

Shai Ben-David and Reba Schuller. 2003. Exploiting task relatedness for multiple task learning. In *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003. Proceedings*, pages 567–580. Springer.

Yaniv Benny, Niv Pekar, and Lior Wolf. 2021. Scale-localized abstract reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12557–12565.

Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schmidt. 2023. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. *arXiv 2308.06595*.

Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. 2022. Iglue: A benchmark for transfer learning across modalities, tasks, and languages. *arXiv 2201.11732*.

Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. 2022. VisualGPT: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18030–18040.

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.

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

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

Poorav Desai, Tanmoy Chakraborty, and Md Shad Akhtar. 2022. Nice perfume. how long did you marinate in it? multimodal sarcasm explanation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10563–10571.

Kshitij Dwivedi and Gemma Roig. 2019. Representation similarity analysis for efficient task taxonomy & transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12387–12396.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369.

Chris Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. 2021. Efficiently identifying task groupings for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 27503–27516.

J.R. Firth. 1957. A synopsis of linguistic theory 1930-1955. *Studies in Linguistic Analysis*, page 1–32.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv 2306.13394*.

Richard L Gorsuch. 2014. *Factor analysis: Classic edition*. Routledge.

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 (CVPR)*, pages 6904–6913.

Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. Do androids laugh at electric sheep? humor “understanding” benchmarks from the new yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, Toronto, Canada. Association for Computational Linguistics.

John L Horn and John J McArdle. 2007. *Understanding human intelligence since Spearman*, pages 205–247.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6700–6709. Computer Vision Foundation / IEEE.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Henry F Kaiser. 1958. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200.

Shankar Kantharaj, Xuan Long Do, Rixie Tiffany Leong, Jia Qing Tan, Enamul Hoque, and Shafiq Joty. 2022. OpenCQA: Open-ended question answering with charts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11817–11837, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *NeurIPS*.

Abhishek Kumar and Hal Daume III. 2012. Learning task grouping and overlap in multi-task learning. *arXiv preprint arXiv:1206.6417*.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

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

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

Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C. H. Hoi. 2022a. Lavis: A library for language-vision intelligence.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023c. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022b. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*.

Zejun Li, Ye Wang, Mengfei Du, Qingwen Liu, Binhao Wu, Jiwen Zhang, Chengxing Zhou, Zhihao Fan, Jie Fu, Jingjing Chen, Xuanjing Huang, and Zhongyu Wei. 2023d. Reform-eval: Evaluating large vision language models via unified re-formulation of task-oriented benchmarks. *arXiv 2310.02569*.

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.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning.

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023d. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *NeurIPS Track on Datasets and Benchmarks*.

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 (CVPR)*, pages 3195–3204.

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.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv 1301.3781*.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE.

OpenAI. 2023a. Chatgpt.

OpenAI. 2023b. Gpt-4.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

John C Raven. 1938. *Raven's Progressive Matrices: Sets A, B, C, D, E.* Australian Council for Educational Research.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 35:25278–25294.

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

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: a dataset for image captioningwith reading comprehension.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *CVPR*, pages 8317–8326.

Xiaozhuang Song, Shun Zheng, Wei Cao, James Yu, and Jiang Bian. 2022. Efficient and effective multi-task grouping via meta learning on task combinations. *Advances in Neural Information Processing Systems*, 35:37647–37659.

C. Spearman. 1904. General intelligence, objectively determined and measured. *American Journal of Psychology*, 15(2):201–292.

Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. 2020. Which tasks should be learned together in multi-task learning? In *International Conference on Machine Learning*, pages 9120–9132. PMLR.

Lin Su, Nan Duan, Edward Cui, Lei Ji, Chenfei Wu, Huaishao Luo, Yongfei Liu, Ming Zhong, Taroon Bharti, and Arun Sacheti. 2021. GEM: A general evaluation benchmark for multimodal tasks. In *Findings of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven C.H. Hoi. 2022. Plug-and-play vqa: Zero-shot vqa by conjoining large pre-trained models with zero training. In *Findings of the Conference on Empirical Methods in Natural Language Processing (Findings of EMNLP)*.

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

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Zhiheng Xi, Rui Zheng, Yuansen Zhang, Xuan-Jing Huang, Zhongyu Wei, Minlong Peng, Mingming Sun, Qi Zhang, and Tao Gui. 2023. Connectivity patterns are task embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11993–12013.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2023. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*.

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

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2.

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

Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. 2018. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722.

Wangchunshu Zhou, Yan Zeng, Shizhe Diao, and Xinsong Zhang. 2022. Vlue: A multi-task benchmark for evaluating vision-language models. In *ICML*.

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

A Data Collection Protocol for OLIVE

OLIVE comprises 9,450 images, 30,120 unique instructions and 47,250 reponses. The examples can be broadly categorized into 4 groups: visual recognition, creative writing, knowledge-based, and elaborated description. Some examples are shown in Figure 1.

In our benchmark OLIVE, we use the text-only version of ChatGPT to generate instructions and outputs for each image. Specifically, we sample images from LAION-Aesthetics (Schuhmann et al., 2022) and use the BLIP-2 (Li et al., 2023c) captioning model to encode the visual information in each image into captions. These generated captions, along with the original LAION captions - which may contain entity-specific knowledge useful for generating varied instruction-output data - are then fed into ChatGPT. Additionally, for each of the aforementioned categories, we manually annotate a few seed examples, and use these as in-context examples to guide ChatGPT.

The instructions and outputs generated by ChatGPT could contain incorrect information due to model hallucination, which undermines their reliability for use as an evaluation benchmark. Recognizing this, we hire an annotation company, Flitto which recruits human annotators to thoroughly inspect and correct any erroneous data. They are task to: 1) ensure that the instructions contain minimal shortcut information, which could enable the model to produce correct outputs without having to understand the image, 2) verify the accuracy of the output and confirm that it is free from harmful content, and 3) fact-check knowledge-based information. This comprehensive review process helps to enhance the overall quality and reliability of the data.

B ChatGPT Prompts for OLIVE

Following (Liu et al., 2023c) and (Taori et al., 2023), we construct prompts for ChatGPT (OpenAI, 2023a) to generate instructions and outputs for different categories: visual recognition, elaborated description, knowledge-based and creative writing. For elaborated description, we randomly sample from a list of instructions which inquire about image description.

Prompt for generating creative writing instructions

You are given several image captions, each describing the same image you are observing. Using your creativity and imagination, think of a new instruction that can be induced from the image captions.

Here are the requirements:
1. Try not to repeat the verb for each instruction to maximize diversity.
2. The language used for the instruction also should be diverse. Either an imperative sentence or a question is permitted.
3. The type of instruction should be diverse.
4. The instruction must not involve counting.
5. Make the instruction challenging by not including the visual content details in the instruction so that one must use the captions to understand the instruction.
6. Replace the name of the object entity with a generic term or category, for example replace bus as this vehicle, dress as this clothing, etc.
7. The format of the instruction should follow the examples shown below. Make sure it is numbered and end with '###'.

Prompt for generating knowledge-based instructions

You are given several image captions, each describing the same image you are observing. Using your creativity and imagination, think of a new instruction that can be induced from the image captions.

Here are the requirements:
1. Try not to repeat the verb for each instruction to maximize diversity.
2. The language used for the instruction also should be diverse. Either an imperative sentence or a question is permitted.
3. The instruction should be diverse and ask a question that requires reasoning, not just simple visual recognition.
4. Given the instruction, one should require first understanding the visual content, then based on the background knowledge or reasoning, either explain why the things are happening that way, or provide guides and help to user's request.
5. Make the instruction challenging by not including the visual content details in the instruction so that the user must use the captions to understand the instruction.
6. Replace the name of the object entity with a generic term or category, for example replace bus as this vehicle, dress as this clothing, etc.
7. The instruction must not involve counting.
8. The format of the instruction should follow the examples shown below. Make sure it is numbered and end with '###'.

Prompt for generating visual recognition instructions

You are given several image captions, each describing the same image you are observing. Using your creativity and imagination, think of a new instruction that can be induced from the image captions.

Here are the requirements:
1. Try not to repeat the verb for each instruction to maximize diversity.
2. The language used for the instruction also should be diverse. Either an imperative sentence or a question is permitted.
3. The instruction should ask about the visual content of the image, including the object types, object actions, object locations, etc. Only include instruction that has definite answers founded in the captions.
4. Include complex instruction that is relevant to the content in the image, for example, asking about background knowledge of the objects in the image, asking to discuss about events happening in the image, etc. Again, do not ask about uncertain details.
5. Make the instruction challenging by not including the visual content details in the instruction so that the one must use the captions to understand the instruction.
6. Replace the name of the object entity with a generic term or category, for example replace bus as this vehicle, dress as this clothing, etc.
7. The instruction must not involve counting.
8. The format of the instruction should follow the examples shown below. Make sure it is numbered and end with '###'.

List of instructions for elaborated description (part 1)

- Provide a vivid description of the image.
- What is a suitable paragraph that describes this image?
- Compose a passage that depicts this image.
- What is this image about?
- What's happening in the scene?
- Can you describe the main features of this image for me?
- What are the key details in this picture?
- Can you elaborate on the elements of the picture provided?
- What do you think is going on in this photo?
- Can you provide a comprehensive description of the image?
- Describe the following image in detail.
- Provide a detailed portrayal of what's captured in this image.
- Offer an intricate description of the image you see.
- Please share a thorough run down of the image that has been presented.
- Could you elaborate on the contents of the displayed image with thoroughness?
- Clarify the contents of the displayed image with elaborate detail.

List of instructions for elaborated description (part 2)

- Can you offer a comprehensive portrayal of the image?
- Could you highlight and elaborate on the details of the image?
- Portray the image with a vivid comprehensive narrative.
- Analyze the image in a descriptive manner.
- Write an well-detailed depiction of the given image.
- How would you describe this photo in great detail?
- Can you give a detailed account of what you see in this image?
- Describe this image using your own words.
- Please describe what you see in the image with as much detail as possible.
- I need you to depict the image with utmost detail.
- Can you describe the image below in exhaustive detail?
- Please provide a complete description of what is shown in the picture.
- I would like you to give a detailed clarification of the contents of the displayed image.
- Could you provide a detailed and comprehensive representation of the image?
- Provide a comprehensive illustration of the image.
- Illustrate the image using a well-detailed description.
- Write a rich narrative for this image.
- Give a thorough description for the given image.
- Write a vivid account of the moment captured in this image.
- Create a narrative that is rich and vivid based on the image presented.

Prompt for generating visual recognition, knowledge-based and creative writing outputs

You are given an instruction and several image captions, each caption describing the same image you are observing. Generate an output resulting from following the instruction.

Here are the requirements:
1. The output is the response to the instruction and the caption.
2. The output must utilize the information in the caption and must not contradict the caption.
3. If the output is unknown without further context, generate "unknown" as the output.
4. When using the information from the caption, directly explain the scene, do not mention that the information source is the caption. Always answer as if you are directly looking at the image.
5. Provide detailed output when answering complex instruction. For example, give detailed examples or reasoning steps to make the content more convincing and well-organized.
6. The format of the output should follow the examples shown below. Make sure it is numbered and end with '###'.

Prompt for generating elaborated description outputs

You are given several image captions, each caption describing the same image you are observing.

Here are the requirements:
1. Generate an output that describes the image in detail.
2. The output must utilize the information in the caption and must not contradict the caption. Do not include description of objects that is not presented in the caption.
3. When using the information from the caption, directly explain the scene, do not mention that the information source is the caption. Always answer as if you are directly looking at the image.
4. The format of the output should follow the examples shown below. Make sure it is numbered and end with '###'.

C Model Details

We experiment with 4 different VLMs as follows:

- BLIP-2 utilizes ViT-G/14 (Fang et al., 2023) as the image encoder and FlanT5$_{\text{XL}}$ (Chung et al., 2022) as the LLM. We initialize BLIP-2 from the pretrained checkpoint and only fine-tune the Q-former parameters. Both the image encoder and the LLM are frozen. The total and trainable parameters are 4B and 187M respectively.

- MiniGPT-4 adopts ViT-G/14 (Fang et al., 2023) as the image encoder and Vicuna$_{\text{7B}}$ (Chiang et al., 2023) as the LLM. It consists of the BLIP-2 Q-former and a linear layer connecting the image encoder and the LLM. The Q-former is initialized from BLIP-2. All parameters are frozen except the linear layer. The total and trainable parameters are 8B and 3M respectively.

- mPLUG-Owl adopts ViT-L/14 (Radford et al., 2021) as the image encoder and LLaMA$_{\text{7B}}$ (Touvron et al., 2023) as the LLM. It consists of a visual abstractor module between the image encoder and the LLM. All parameters are frozen except LoRA (Hu et al., 2022) parameters on the LLM. The total and trainable parameters are 7B and 4M respectively.

- LLaVA adopts ViT-L/14 (Radford et al., 2021) as the image encoder and $\text{LLaMA}_{7\text{B}}$ (Touvron et al., 2023) as the LLM. It consists of a linear layer in between the image encoder and the LLM. All parameters are frozen except the linear layer and LoRA (Hu et al., 2022) parameters on the LLM. The total and trainable parameters are 7B and 164M respectively.

D Additional Hyperparameters

We individually finetune models for each task using datasets in the instruction format. Only one instruction template is used per task, as preliminary experiments show using multiple templates per task degrades performance.

For all experiments using the same model architecture, we keep the hyperparameters constant. We set the training iteration to 10K steps. The batch size for BLIP-2 is 192 and 128 for the other three models. For BLIP-2, MiniGPT-4 and mPLUG-Owl, we train the model using AdamW optimizer with a weight decay of 0.05. The learning rate is linearly increased from $1\text{e}{-8}$ to $1\text{e}{-5}$ in the first 200 steps and then cosine decayed to 0. For LLaVA, we use a weight decay of 0. The learning rate is linearly rises from 0 to $2\text{e}{-5}$ across the initial 200 steps and then cosine decay to 0.

We output model performances at intervals of 1,000 iterations and select the best checkpoint using the validation set for evaluation.

All experiments are performed on a machine with 8 or 16 Nvidia A100 GPUs. On average, each experiment involves around 2 hours of training and another 2 hours of evaluation. We utilize LAVIS (Li et al., 2022a) library for training of BLIP-2, MiniGPT-4 and mPLUG-Owl. For LLaVA, we utilize LLaVA original author's codebase for training. All evaluations are performed on LAVIS.

E Complete Results

In this section, we show all the experimental results from all four models. Tables 4-7 show the raw transfer learning performance, where rows denote the source tasks and columns are target tasks. Tables 8-11 show the normalized performance, where the rows (source tasks) are sorted in a descending order of average performance.

Among the source tasks, LLaVA Conversation shows strong transfer to OLIVE for BLIP-2 and LLaVA. The relatively good transferability could attribute to the fact that LLaVA Conversation and OLIVE share some similarities in data distribution since the instruction-response pairs are generated using OpenAI GPT models (OpenAI, 2023a,b). However, the key difference is that OLIVE is inspected by human annotators to rectify erroneous data, while LLaVA Conversation does not undergo this correction process.

Source Task	Dataset Size	COCO Caption	Flickr 30k	Text Caps	VQAv2 G	VQAv2 MC	OK-VQA G	OK-VQA MC	A-OKVQA G	A-OKVQA MC	Science QA	GQA G	GQA MC	Icon QA	VSR	CLEVR G	CLEVR MC	RAVEN-FAIR	Text VQA G	Text VQA MC	OCR-VQA G	OCR-VQA MC	Open CQA	Chart QA G	Chart QA MC	HM	NY Explain	NY Rank	MORE	OLIVE
Zero-shot	-	128.8	79.2	71.4	63.0	64.7	40.9	59.2	43.6	70.2	69.6	43.9	46.8	48.0	65.7	33.9	36.9	12.4	26.6	62.7	36.9	72.6	6.9	9.1	51.1	55.0	8.4	53.5	14.4	5.2
COCO Caption	567 K	140.9	83.0	75.1	32.5	63.3	23.8	56.1	27.2	70.7	68.8	33.1	45.9	51.9	66.2	31.0	34.9	12.4	18.3	57.7	37.6	71.5	10.6	8.0	46.4	55.6	9.5	53.4	11.1	7.9
Flickr30k	145 K	112.7	99.1	80.0	57.9	65.2	35.3	57.7	42.5	72.8	68.1	40.3	46.5	51.1	66.2	32.9	35.5	12.4	25.7	62.2	37.5	72.8	10.4	8.4	47.0	56.6	9.8	52.8	12.8	6.0
Web CapFilt	23,147 K	134.2	81.4	78.9	60.7	65.0	38.5	57.6	43.5	71.5	68.6	42.4	47.1	48.9	63.9	31.6	33.4	12.4	21.9	62.0	33.7	71.7	9.5	7.7	45.3	56.2	9.3	52.5	14.7	9.2
TextCaps	549 K	65.5	46.6	106.0	36.7	62.4	31.3	56.2	37.1	70.7	68.2	34.7	44.1	52.2	61.5	31.1	35.9	12.4	26.0	66.3	36.8	72.6	10.4	8.6	45.4	56.0	10.6	51.8	13.3	7.7
VQAv2	444 K	74.3	46.9	45.3	72.4	71.2	48.2	61.6	53.9	75.5	68.4	47.9	50.3	55.1	69.1	36.8	39.6	12.5	31.4	66.8	37.9	70.8	1.3	10.2	43.7	58.0	4.2	54.9	9.5	2.3
OK-VQA	9 K	75.8	48.7	51.6	57.7	65.2	51.5	64.4	44.1	70.6	66.4	40.5	46.8	47.0	62.8	29.8	36.4	12.4	25.0	61.4	36.3	69.3	0.8	9.8	43.6	56.6	4.0	54.5	10.5	2.2
A-OKVQA	17 K	43.5	28.7	30.3	62.3	66.7	48.8	62.4	55.3	74.7	68.1	41.9	46.7	49.0	64.4	34.1	38.6	12.5	25.3	62.0	36.5	68.4	0.7	8.9	40.1	54.0	2.5	53.2	8.4	2.2
A-OKVQA (MC)	17 K	115.8	78.0	69.3	63.7	66.1	41.8	58.4	47.2	76.4	69.7	43.2	47.3	48.2	63.6	36.0	37.3	12.4	27.0	62.8	37.2	74.3	6.3	9.6	53.2	56.6	8.1	54.5	12.8	4.4
ScienceQA	6 K	120.7	74.3	69.3	62.2	63.1	39.9	59.0	43.8	69.4	91.5	42.9	47.8	46.6	62.2	32.6	31.9	12.2	25.4	60.5	33.0	71.5	3.2	4.9	43.8	56.0	5.0	55.2	5.1	3.9
GQA	943 K	25.1	18.6	17.2	60.2	68.1	33.8	60.3	40.1	68.9	68.9	55.3	57.0	47.1	62.0	31.7	40.7	12.5	16.1	57.1	34.4	69.3	0.4	6.8	50.1	54.8	1.4	54.8	1.8	1.7
IconQA	19 K	122.6	77.2	68.0	61.3	64.2	34.2	56.9	38.5	67.9	67.6	43.5	47.7	75.3	63.8	25.1	33.9	12.8	23.6	59.6	34.8	64.3	2.5	6.5	36.2	52.4	3.6	54.5	5.6	3.6
VSR	3 K	107.7	65.4	64.8	13.2	46.8	3.6	44.1	3.2	59.7	61.2	8.4	42.1	45.1	65.0	0.2	33.4	12.5	4.5	42.2	13.3	59.4	2.2	0.4	19.4	54.4	0.5	52.9	8.9	2.0
TextVQA	35 K	1.4	0.5	20.2	49.3	59.1	33.9	57.4	36.0	65.2	68.0	36.3	44.9	46.0	55.2	31.3	36.3	12.3	36.0	70.1	37.7	72.3	1.9	10.8	47.4	55.8	6.0	53.9	13.7	2.3
OCR-VQA	802 K	113.5	70.9	55.0	35.6	50.6	8.9	48.8	13.2	54.5	63.5	28.3	41.2	42.9	58.4	21.4	33.6	12.5	11.2	40.9	56.9	67.3	0.0	2.5	33.3	53.8	0.1	51.9	1.4	1.8
OpenCQA	6 K	94.0	72.0	61.8	57.9	63.6	33.3	55.2	38.7	69.3	65.5	42.1	47.0	46.6	61.4	25.0	41.8	12.5	16.7	56.6	16.0	53.5	34.9	0.6	40.9	50.8	13.8	50.7	14.9	10.2
HM	9 K	116.0	70.5	64.0	32.1	63.7	0.6	57.1	1.3	66.7	67.1	20.9	44.2	47.6	64.2	21.0	37.5	12.2	4.3	55.4	30.3	58.4	0.0	1.6	36.7	68.4	0.4	53.8	1.7	2.4
OLIVE	7 K	13.5	16.5	17.6	45.4	63.2	16.4	56.3	20.0	67.5	67.7	38.4	46.5	49.8	65.6	26.8	39.4	12.5	13.0	61.8	31.5	71.4	11.0	4.4	48.7	51.0	15.9	53.1	13.0	38.1
LLaVA Conversation	57 K	59.5	40.6	42.1	6.9	57.1	0.9	47.2	1.2	62.3	61.8	5.9	42.5	46.8	55.1	3.9	23.4	12.5	2.3	50.3	3.4	46.8	26.8	0.8	34.8	50.2	15.1	48.5	11.7	38.7
LLaVA Reasoning	77 K	9.4	11.0	11.6	0.0	56.1	0.0	48.0	0.0	62.0	66.3	0.0	44.4	45.3	59.6	0.3	41.3	12.4	0.0	42.2	0.0	48.6	4.7	0.2	40.9	50.2	15.6	52.2	11.7	8.8
LLaVA Description	23 K	13.1	15.9	14.7	0.0	19.7	0.0	28.9	0.0	47.0	62.2	0.0	16.9	39.8	47.7	0.0	37.3	12.5	0.0	21.0	0.0	23.3	3.5	0.0	17.4	46.4	14.5	47.0	7.9	6.8
VQAv2 QG	444 K	31.2	12.3	20.1	37.2	59.5	19.2	55.2	25.0	66.9	65.2	34.0	45.4	45.6	62.1	26.5	39.4	12.6	15.9	61.5	30.6	38.9	10.2	2.0	37.3	51.4	7.1	51.8	10.5	8.1
OK-VQA QG	9 K	18.0	13.6	19.6	0.5	44.2	1.2	45.7	1.7	54.8	61.4	0.9	34.0	43.7	55.6	0.1	36.9	12.5	1.6	44.9	23.4	32.7	6.8	1.8	32.5	50.8	6.4	49.7	10.0	7.1
A-OKVQA QG	17 K	33.3	19.8	32.2	10.7	57.0	12.8	52.1	16.2	66.7	65.6	14.4	44.1	49.0	63.0	9.9	34.3	12.5	9.9	52.8	29.5	39.8	7.6	3.0	36.0	56.0	8.5	48.5	11.5	6.0

Table 4: Unnormalized transfer learning performance of BLIP-2. Higher values indicate better performance. QG denotes question generation, MC denotes multiple-choice and G denotes open-ended generation. The color scale is normalized along each column. The colors represent values in descending order: green, yellow, orange and red.

Source Task	Dataset Size	COCO Caption	Flickr 30k	Text Caps	VQAv2		OK-VQA		A-OKVQA		Science QA	GQA		Icon QA	VSR	CLEVR		RAVEN-FAIR	Text VQA		OCR-VQA		Open CQA	Chart QA		HM	NY Explain	NY Rank	MORE	OLIVE
					G	MC	G	MC	G	MC		G	MC			G	MC		G	MC	G	MC		G	MC					
Zero-shot	-	9.5	14.2	13.3	35.7	44.5	36.4	40.9	36.5	33.3	46.8	27.1	34.2	39.2	58.7	25.5	35.8	12.5	17.9	45.8	14.8	27.3	2.8	3.7	49.0	53.6	8.8	52.3	5.6	29.6
COCO Caption	567 K	133.9	75.5	65.2	4.1	50.6	9.0	47.0	6.8	38.1	45.9	2.2	33.7	41.9	55.6	1.8	36.7	12.4	3.5	52.8	1.7	36.2	15.2	0.8	29.6	55.6	10.5	49.7	12.0	16.4
Flickr30k	145 K	96.1	92.2	71.4	39.2	50.2	34.8	47.6	36.2	45.7	52.1	32.2	37.9	44.1	59.2	31.3	36.1	12.4	25.2	58.6	24.2	38.7	13.3	6.5	40.6	59.4	10.8	52.3	11.3	19.8
Web CapFilt	23,147 K	113.7	76.9	75.7	16.1	47.3	21.0	48.4	24.2	53.7	51.0	11.8	33.9	44.1	52.4	9.4	31.8	12.5	10.9	54.5	2.3	24.7	9.2	0.3	17.4	49.0	8.3	48.7	12.8	10.4
TextCaps	549 K	56.2	42.0	111.8	23.2	33.8	24.1	47.1	26.4	43.6	45.0	17.6	25.3	41.6	54.8	22.0	27.6	12.4	15.9	55.3	22.9	33.9	20.4	3.0	23.8	52.6	11.8	49.8	15.0	17.3
VQAv2	444 K	57.9	38.5	32.5	72.5	67.6	55.5	59.6	58.6	65.2	57.6	47.4	45.2	48.0	62.5	35.9	37.1	12.4	38.3	65.1	38.5	33.4	6.7	12.0	29.0	59.2	1.2	51.0	2.1	2.0
OK-VQA	9 K	51.2	34.0	29.3	52.2	54.3	56.3	58.5	46.0	57.2	57.6	35.2	40.8	47.4	48.5	22.1	35.5	12.5	27.5	61.6	23.0	39.6	4.9	9.8	37.6	56.4	4.0	54.8	8.8	2.0
A-OKVQA	17 K	41.5	27.5	24.0	60.6	62.9	53.1	61.4	57.1	67.5	57.0	40.5	43.5	48.1	48.5	28.5	37.6	13.0	30.1	63.8	27.3	22.3	0.9	9.4	33.7	56.4	0.8	52.3	3.9	1.8
A-OKVQA (MC)	17 K	25.9	25.5	24.7	47.3	58.4	43.9	59.6	45.2	76.1	63.8	29.1	42.2	46.1	49.7	23.1	36.3	12.6	28.0	70.2	24.0	67.2	10.7	7.7	54.8	55.6	9.3	52.9	6.4	13.2
ScienceQA	6 K	17.8	18.5	20.1	48.7	53.1	43.6	54.1	44.9	62.4	79.1	34.8	38.2	46.2	53.0	27.4	31.9	12.8	25.4	62.6	24.1	60.6	8.3	9.2	37.3	53.4	9.7	55.2	6.2	32.3
GQA	943 K	46.0	27.7	20.2	58.1	62.5	37.8	55.8	41.9	63.3	55.9	52.8	50.3	45.3	61.1	31.1	37.7	12.4	17.4	51.4	29.1	41.9	0.5	4.4	35.2	57.6	9.9	54.0	6.5	2.1
IconQA	19 K	15.6	17.0	20.2	46.1	55.1	42.5	49.1	43.1	54.1	56.1	33.8	39.0	65.9	53.4	27.4	33.0	12.2	24.5	56.6	25.1	61.1	8.9	8.4	34.2	54.8	10.0	53.9	6.5	27.4
VSR	3 K	9.9	14.0	14.3	51.4	47.9	38.9	42.3	41.1	34.8	47.4	38.6	35.9	40.9	64.3	29.5	35.8	12.5	21.4	46.7	31.8	43.9	3.8	4.3	51.7	50.0	8.7	52.0	5.8	30.0
TextVQA	35 K	80.6	57.7	61.0	58.3	57.4	44.5	54.7	45.8	62.6	59.1	39.3	40.4	45.6	57.4	33.7	34.3	12.5	43.5	66.9	39.2	45.9	4.7	11.9	33.0	57.8	4.3	48.8	5.2	3.1
OCR-VQA	802 K	58.6	44.5	53.8	54.6	44.7	41.0	43.5	43.8	51.1	56.4	40.7	40.2	42.1	54.9	33.1	21.2	12.5	31.1	44.9	60.6	75.8	8.0	9.8	40.8	58.2	10.6	50.2	7.9	3.8
OpenCQA	6 K	28.3	22.8	27.7	38.1	42.4	34.0	45.5	32.2	44.6	55.7	27.2	33.1	44.0	53.3	19.5	33.8	12.4	20.6	49.9	18.3	36.0	24.0	3.9	41.2	50.6	11.6	51.0	11.9	20.4
HM	9 K	9.9	14.2	13.9	43.8	44.7	35.3	40.0	34.2	32.8	43.6	32.4	33.4	41.7	52.0	25.4	34.3	12.4	17.1	39.4	21.8	39.0	3.3	3.7	46.0	55.8	8.7	50.5	5.5	31.9
OLIVE	7 K	8.6	12.0	14.0	28.5	41.3	16.2	38.6	16.2	34.7	45.9	20.6	32.8	38.5	49.8	11.6	26.3	12.2	12.3	44.2	10.5	28.5	7.2	0.6	37.2	50.2	10.1	51.8	6.5	30.6
LLaVA Conversation	57 K	20.3	15.9	21.1	3.0	41.7	12.6	43.4	8.2	35.3	50.3	1.7	29.1	38.8	55.6	0.8	28.3	12.4	6.4	44.6	0.4	27.7	9.7	0.7	37.4	58.6	9.0	48.7	5.6	31.8
LLaVA Reasoning	77 K	10.5	13.7	11.0	41.2	43.0	29.4	39.8	29.6	30.7	44.7	31.5	27.3	35.7	52.3	21.1	34.5	12.4	18.8	40.8	0.0	19.2	2.1	0.0	43.5	54.8	9.1	50.3	6.0	3.4
LLaVA Description	23 K	8.5	11.4	10.8	0.0	34.5	0.0	33.3	0.0	24.4	34.3	0.0	24.4	35.1	51.7	0.0	31.0	12.0	0.0	33.8	0.0	34.1	1.8	0.0	29.3	50.6	9.7	50.5	5.6	2.1
VQAv2 QG	444 K	58.5	47.5	25.4	16.0	47.7	23.3	42.6	24.6	46.0	47.6	10.7	39.4	42.1	52.0	8.1	34.4	12.4	12.2	51.8	6.9	30.2	8.7	1.6	35.5	55.4	11.2	50.3	10.7	8.3
OK-VQA QG	9 K	31.4	25.0	17.7	12.9	39.7	18.6	39.0	20.3	36.9	42.7	7.8	28.2	43.2	49.2	11.6	32.7	12.2	8.0	41.1	3.7	23.4	8.4	0.9	36.6	50.4	12.6	52.1	11.2	11.8
A-OKVQA QG	17 K	19.3	16.2	15.4	12.0	41.9	12.5	40.1	15.1	30.7	41.5	7.7	30.6	40.1	53.8	10.5	34.2	12.7	6.5	38.8	3.1	22.4	5.3	1.2	31.0	46.2	10.4	49.7	10.4	7.0

Table 5: Unnormalized transfer learning performance of LLaVA. Higher values indicate better performance. QG denotes question generation, MC denotes multiple-choice and G denotes open-ended generation. The color scale is normalized along each column. The colors represent values in descending order: green, yellow, orange and red.

Source Task	Dataset Size	COCO Caption	Flickr 30k	Text Caps	VQAv2		OK-VQA		A-OKVQA		Science QA	GQA		Icon QA	VSR	CLEVR		RAVEN-FAIR	Text VQA		OCR-VQA		Open CQA	Chart QA		HM	NY Explain	NY Rank	MORE	OLIVE
					G	MC	G	MC	G	MC		G	MC			G	MC		G	MC	G	MC		G	MC					
Zero-shot	-	13.3	14.5	12.4	0.2	48.6	0.1	39.5	0.0	33.4	43.6	0.0	35.8	41.2	52.7	0.0	34.1	12.4	0.6	47.7	0.0	48.0	15.8	1.9	41.8	49.8	11.3	49.0	5.8	1.6
COCO Caption	567 K	136.1	80.4	68.1	0.0	44.9	0.0	39.9	0.0	32.4	44.3	0.0	33.9	43.6	51.4	0.0	35.0	12.5	0.4	45.5	0.3	47.3	10.4	1.7	46.4	50.4	9.7	47.7	10.0	9.9
Flickr30k	145 K	109.4	92.2	67.0	1.0	51.3	1.4	46.6	0.8	49.0	45.7	0.7	37.5	42.2	59.1	0.1	33.7	12.7	3.3	50.3	0.6	52.2	12.8	5.2	42.9	49.8	10.8	48.0	10.3	11.0
Web CapFilt	23,147 K	127.4	77.3	71.2	1.8	53.5	1.6	48.9	1.0	52.4	47.5	1.9	39.6	41.2	50.2	0.0	34.0	12.5	2.1	50.5	0.9	30.9	8.2	0.0	30.6	48.4	8.0	48.1	9.6	9.4
TextCaps	549 K	91.3	62.0	99.7	0.7	48.3	1.0	44.2	0.8	42.4	43.7	0.4	36.8	39.3	51.4	0.0	34.1	13.0	1.1	51.8	0.1	27.3	12.7	0.0	19.4	49.6	10.0	46.8	14.0	9.8
VQAv2	444 K	31.7	22.9	27.2	68.7	66.3	50.2	58.0	52.5	64.7	53.3	44.8	44.0	43.3	63.3	34.8	35.0	12.4	27.0	57.5	31.5	57.0	3.8	9.3	39.0	51.8	6.9	48.1	6.3	2.5
OK-VQA	9 K	18.2	11.3	20.5	56.2	55.3	54.1	59.5	44.6	59.7	52.2	36.6	36.0	40.9	60.9	33.6	32.9	12.0	25.4	56.7	15.1	50.9	2.0	8.6	40.6	53.0	8.1	50.1	8.7	1.9
A-OKVQA	17 K	28.0	18.8	22.5	59.7	58.8	50.7	61.4	55.1	64.1	53.0	39.0	38.2	44.2	62.4	31.7	33.3	12.4	26.0	55.4	20.2	56.1	2.3	9.0	40.0	52.6	9.3	48.2	8.2	2.3
A-OKVQA (MC)	17 K	39.6	27.2	34.7	58.2	57.7	41.7	58.8	42.9	73.2	57.1	35.1	35.0	45.3	54.1	5.7	33.5	12.4	19.3	61.1	2.0	55.4	12.7	6.6	42.9	51.2	12.7	49.5	12.7	14.3
ScienceQA	6 K	60.0	45.5	40.2	48.7	45.6	32.0	49.6	28.8	54.2	77.8	32.9	32.5	44.0	52.5	25.1	31.6	13.1	17.8	56.8	3.6	51.0	4.6	5.6	33.7	51.0	13.6	50.8	13.7	1.1
GQA	943 K	16.2	8.7	21.1	60.5	62.6	38.0	55.0	43.3	59.6	48.6	50.3	48.1	43.3	53.2	36.2	34.2	12.4	19.1	49.0	25.2	52.4	2.9	8.2	38.4	51.4	9.9	47.7	7.9	2.3
IconQA	19 K	55.2	36.8	37.4	17.2	47.2	8.4	40.0	8.4	42.1	44.9	10.7	35.3	68.5	54.0	18.6	36.2	13.0	6.9	43.8	6.5	29.0	4.7	6.1	33.0	48.2	12.6	50.5	11.5	18.0
VSR	3 K	50.0	37.0	22.9	3.9	53.4	2.1	40.3	2.2	44.2	45.4	3.3	35.9	44.6	63.5	0.0	38.7	12.4	1.3	50.2	0.1	51.3	19.8	2.8	41.1	51.6	12.3	48.0	9.6	6.8
TextVQA	35 K	2.2	0.5	18.8	33.7	55.0	32.4	50.4	35.1	51.4	54.6	25.1	39.6	43.7	60.0	31.0	31.8	12.5	33.1	59.0	32.4	36.0	0.4	10.1	32.9	52.2	4.0	48.5	3.7	2.3
OCR-VQA	802 K	92.4	56.3	48.5	16.8	57.1	13.1	45.8	12.2	44.5	52.6	9.0	40.9	41.6	62.4	5.9	35.8	12.4	16.5	49.6	53.1	70.1	8.1	8.3	41.5	50.2	12.3	48.4	10.9	7.1
OpenCQA	6 K	28.5	20.5	21.8	0.2	50.9	0.2	42.0	0.1	38.1	51.0	0.0	38.2	41.1	52.5	0.0	35.1	12.4	0.8	53.1	0.2	52.0	29.8	0.0	49.0	53.0	12.2	49.9	10.3	18.3
HM	9 K	103.1	65.7	50.1	48.1	54.2	20.3	41.0	19.0	39.7	45.8	31.9	38.4	41.7	56.1	22.8	34.1	12.5	12.8	46.2	22.0	47.4	7.6	8.4	37.4	70.6	10.9	47.3	12.3	14.2
OLIVE	7 K	14.7	16.7	19.8	0.0	44.3	0.1	38.8	0.0	34.1	44.2	0.0	33.7	42.3	58.2	0.0	37.5	12.4	0.5	45.9	0.1	49.1	15.4	0.4	37.9	52.4	10.7	50.2	7.8	34.1
LLaVA Conversation	57 K	29.9	23.1	28.4	0.0	24.3	0.0	35.5	0.0	28.0	40.8	0.0	16.6	41.5	56.5	0.0	39.1	12.4	0.1	40.5	0.0	49.8	16.9	0.4	44.3	54.6	10.9	48.8	8.7	1.2
LLaVA Reasoning	77 K	12.0	16.1	15.0	0.0	41.2	0.0	38.5	0.0	30.3	40.1	0.0	30.1	40.6	56.5	0.0	38.3	12.4	0.2	40.7	0.0	45.3	9.4	0.3	38.3	49.6	9.7	49.9	6.4	0.9
LLaVA Description	23 K	10.0	13.8	13.4	0.0	36.7	0.0	33.7	0.0	19.8	37.4	0.0	27.8	38.6	48.1	0.0	34.4	12.7	0.0	32.0	0.0	39.4	9.0	0.0	36.2	48.6	10.5	48.2	7.2	2.7
VQAv2 QG	444 K	85.3	52.4	50.8	13.1	43.8	13.8	38.2	13.7	30.5	46.1	6.5	32.3	43.3	52.0	2.5	34.1	12.4	6.8	44.9	0.9	53.8	14.0	5.0	41.0	50.2	12.2	49.0	13.2	4.1
OK-VQA QG	9 K	27.7	21.8	32.0	1.1	44.2	1.9	36.8	1.5	33.2	44.8	0.4	33.0	44.1	51.7	0.0	36.2	12.4	2.6	48.7	0.2	53.4	15.8	4.8	40.0	53.8	11.9	49.3	11.6	15.1
A-OKVQA QG	17 K	40.1	25.6	37.5	8.0	45.6	0.7	39.8	1.0	32.1	45.4	4.7	33.1	42.7	56.5	0.2	36.6	12.4	3.1	42.1	0.4	54.1	11.8	5.8	39.8	51.4	12.1	49.3	11.9	15.4

Table 6: Unnormalized transfer learning performance of MiniGPT-4. Higher values indicate better performance. QG denotes question generation, MC denotes multiple-choice and G denotes open-ended generation. The color scale is normalized along each column. The colors represent values in descending order: green, yellow, orange and red.

Source Task	Dataset Size	COCO Caption	Flickr 30k	Text Caps	VQAv2 G	VQAv2 MC	OK-VQA G	OK-VQA MC	A-OKVQA G	A-OKVQA MC	Science QA	GQA G	GQA MC	Icon QA	VSR	CLEVR G	CLEVR MC	RAVEN-FAIR	Text VQA G	Text VQA MC	OCR-VQA G	OCR-VQA MC	Open CQA	Chart QA G	Chart QA MC	HM	NY Explain	NY Rank	MORE	OLIVE
Zero-shot	-	16.6	15.2	26.2	0.0	48.0	0.0	38.7	0.0	35.5	43.4	0.0	35.5	40.9	55.9	0.0	31.4	12.4	0.0	47.9	0.0	51.8	0.6	0.0	42.6	50.2	10.8	50.6	8.0	4.4
COCO Caption	567 K	123.9	75.4	69.8	0.0	49.0	0.1	40.1	0.1	43.8	41.6	0.0	36.2	41.5	52.0	0.0	35.5	12.4	0.6	58.8	2.4	44.1	14.8	0.2	52.5	50.0	11.0	49.1	12.8	26.9
Flickr30k	145 K	96.7	83.2	73.3	3.2	52.9	6.8	44.2	7.5	48.4	43.8	2.8	37.0	44.2	50.5	0.1	34.0	12.6	5.4	58.3	5.8	48.4	16.3	1.7	44.6	51.8	11.3	49.2	12.4	27.6
Web CapFilt	23,147 K	116.0	76.4	83.3	11.6	53.7	16.4	47.3	19.1	48.5	43.7	9.6	40.5	36.4	48.2	2.0	32.2	12.1	9.3	54.9	4.7	27.0	11.7	1.9	16.5	50.8	9.4	47.6	14.7	13.9
TextCaps	549 K	71.5	50.7	109.4	4.9	44.3	8.5	41.9	8.2	39.0	43.1	3.1	35.1	38.8	52.9	0.1	35.7	12.8	10.9	53.7	6.0	38.7	16.5	1.0	44.7	50.0	11.4	49.2	16.2	24.7
VQAv2	444 K	61.0	40.0	39.0	67.5	65.3	54.2	61.8	56.7	66.8	58.1	44.0	43.3	45.8	62.9	39.7	38.6	12.7	37.6	68.3	42.0	48.2	0.3	12.4	24.5	54.8	4.4	51.3	5.0	2.1
OK-VQA	9 K	54.0	33.2	33.8	57.7	58.2	59.4	63.0	46.9	60.9	56.8	38.2	40.3	44.9	49.8	31.3	37.6	12.3	32.3	62.5	38.3	32.7	0.2	12.0	29.1	50.0	2.6	51.9	4.4	1.6
A-OKVQA	17 K	50.6	34.1	27.6	62.2	63.4	54.9	64.5	58.3	66.0	57.7	41.9	41.9	47.0	49.4	38.0	39.7	12.4	33.9	66.5	39.8	50.3	0.2	10.3	39.5	50.0	2.0	53.2	4.5	1.8
A-OKVQA (MC)	17 K	88.9	58.7	50.1	56.4	53.3	43.8	62.0	44.7	73.7	62.8	32.5	32.8	46.8	58.9	22.7	38.0	12.4	30.6	68.4	29.9	66.7	3.8	11.0	46.0	50.2	11.7	51.8	9.7	13.2
ScienceQA	6 K	69.4	49.8	53.9	31.1	57.6	30.5	55.1	28.5	68.9	79.3	26.9	37.5	44.8	58.8	20.9	36.8	12.8	27.4	64.0	32.4	65.3	8.5	6.3	28.4	50.0	11.7	50.5	12.5	4.8
GQA	943 K	54.5	35.7	24.0	58.8	63.1	41.5	62.6	44.7	64.5	54.7	49.7	48.2	44.2	53.5	33.4	36.8	13.0	20.7	56.8	32.2	43.3	0.0	7.7	32.3	50.0	3.2	52.7	2.3	1.5
IconQA	19 K	60.2	43.4	46.4	47.4	56.8	31.7	50.8	31.2	56.9	57.4	34.3	35.7	61.0	58.2	35.0	32.6	12.9	31.4	50.4	39.5	44.4	9.3	12.2	19.3	54.0	11.4	50.6	9.1	11.2
VSR	3 K	56.5	43.0	45.2	39.9	58.9	17.1	50.9	14.8	53.5	50.2	28.9	40.6	44.7	65.2	13.5	33.5	12.3	22.1	57.2	22.0	72.0	14.6	3.3	39.2	50.4	10.8	51.3	8.6	11.6
TextVQA	35 K	92.1	63.8	54.5	48.9	55.1	37.7	56.2	37.5	57.7	51.9	32.1	39.0	43.9	57.6	31.4	35.4	12.2	42.8	59.4	41.9	29.5	1.1	12.5	31.5	55.4	3.0	48.2	2.6	2.2
OCR-VQA	802 K	99.6	69.2	59.8	52.0	59.8	33.6	51.2	39.4	57.3	54.3	36.9	41.2	42.5	54.1	38.8	35.1	12.8	29.7	48.6	60.1	75.4	2.1	11.9	43.8	61.6	11.3	52.1	1.6	2.2
OpenCQA	6 K	64.7	46.5	60.2	3.9	50.3	8.3	46.3	5.1	43.4	52.8	1.8	36.4	44.8	50.2	0.1	33.0	12.0	14.0	53.2	5.9	58.0	29.7	1.6	41.4	55.4	12.2	50.2	14.7	27.5
HM	9 K	48.6	32.5	47.2	31.3	52.9	2.9	42.3	2.7	41.8	42.9	21.4	37.0	45.2	48.8	20.8	32.7	12.5	10.1	43.9	34.7	43.5	15.9	2.8	33.3	71.2	11.1	49.9	10.2	23.2
OLIVE	7 K	14.9	14.8	23.1	0.0	47.4	0.0	42.1	0.0	38.4	40.9	0.0	36.3	42.4	56.0	0.0	23.2	12.2	0.1	43.6	0.1	51.2	15.8	0.0	37.4	50.0	10.7	48.8	8.6	40.3
LLaVA Conversation	57 K	54.4	37.1	41.1	0.0	47.8	0.0	37.8	0.0	34.0	41.2	0.0	33.6	39.9	49.2	0.0	22.7	12.4	0.0	45.5	0.0	49.9	15.2	0.0	39.6	50.0	10.3	49.5	7.6	5.2
LLaVA Reasoning	77 K	11.5	13.7	16.9	0.0	43.4	0.0	32.9	0.0	30.7	37.3	0.0	32.9	39.2	55.3	0.0	31.9	12.4	0.0	36.6	0.0	44.9	5.1	0.0	37.6	50.0	9.7	49.4	6.6	14.0
LLaVA Description	23 K	9.2	12.0	14.2	0.0	42.1	0.0	37.9	0.0	29.6	37.8	0.0	32.7	37.8	51.5	0.0	30.2	12.2	0.0	38.0	0.1	35.5	4.6	0.0	45.7	50.0	10.6	48.8	7.9	4.0
VQAv2 QG	444 K	96.6	64.9	43.1	3.1	39.3	0.4	34.4	0.1	37.1	38.4	1.7	31.0	39.5	51.5	1.1	37.5	12.4	1.2	46.1	7.8	28.2	10.1	0.8	40.0	50.0	11.1	50.9	11.4	18.3
OK-VQA QG	9 K	94.2	62.3	47.8	0.1	31.5	0.1	42.3	0.0	36.4	38.9	0.0	22.9	40.4	48.5	0.0	33.0	12.4	0.5	43.3	1.0	21.9	11.9	0.3	32.2	51.4	4.7	49.0	10.4	4.8
A-OKVQA QG	17 K	71.1	49.6	43.5	3.0	37.3	0.0	43.9	0.1	36.8	41.2	0.5	31.1	39.7	51.6	0.2	38.7	12.2	1.1	46.9	5.4	39.3	11.1	0.2	34.8	53.6	9.1	49.3	11.2	15.6

Table 7: Unnormalized transfer learning performance of mPLUG-Owl. Higher values indicate better performance. QG denotes question generation, MC denotes multiple-choice and G denotes open-ended generation. The color scale is normalized along each column. The colors represent values in descending order: green, yellow, orange and red.

Source Task	Dataset Size	AHP Ranking Score	Target Task																												
			COCO Caption	Flickr 30k	Text Caps	VQAv2		OK-VQA		A-OKVQA		Science QA	GQA		Icon QA	VSR	CLEVR		RAVEN-FAIR	Text VQA		OCR-VQA		Open CQA	Chart QA		HM	NY Explain	NY Rank	MORE	OLIVE
						G	MC	G	MC	G	MC		G	MC			G	MC		G	MC	G	MC		G	MC					
A-OKVQA (MC)	17 K	10.3	-10.6	-0.6	-0.6	0.7	2.2	0.8	-1.7	3.2	10.0	0.0	-0.7	0.5	0.1	-6.3	7.1	0.9	-0.8	0.4	0.1	0.2	10.0	-0.2	2.8	10.0	1.2	-0.4	5.9	-28.7	-0.3
VQAv2	444 K	9.2	-44.8	-16.3	-7.5	10.0	10.0	6.8	4.5	8.8	8.6	-0.6	3.5	3.4	2.6	10.0	10.0	5.5	1.9	5.2	5.6	0.5	-11.2	-2.0	6.3	-34.1	2.2	-5.6	8.2	-90.0	-0.9
Web CapFilt	23,147 K	8.0	4.4	1.1	2.2	-2.4	0.5	-2.2	-3.2	0.0	2.1	-0.5	-1.4	0.3	0.3	-5.4	-7.5	-7.2	0.0	-5.0	-1.0	-1.6	-5.8	0.9	-7.9	-26.7	0.9	1.1	-5.9	6.1	1.2
Flickr30k	145 K	6.2	-13.2	10.0	2.5	-5.4	0.8	-5.3	-2.9	-0.9	4.1	-0.7	-3.2	-0.3	1.1	1.5	-3.2	-2.9	0.0	-0.9	-0.7	0.3	1.3	1.3	-3.7	-18.9	1.2	1.9	-4.1	-29.8	0.2
ScienceQA	6 K	5.8	-6.6	-2.5	-0.6	-0.8	-2.4	-1.0	-0.4	0.2	-1.3	10.0	-0.9	1.0	-0.5	-10.5	-4.3	-10.3	-4.3	-1.3	-2.9	-1.9	-6.7	-1.3	-24.2	-33.5	0.7	-4.5	10.0	-169.7	-0.4
A-OKVQA	17 K	5.7	-70.1	-25.5	-11.9	-0.7	3.0	7.4	6.1	10.0	7.2	-0.7	-1.8	-0.1	0.4	-3.9	0.8	3.5	1.9	-1.3	-1.0	-0.2	-25.5	-2.2	-1.2	-50.7	-0.7	-7.9	-1.6	-110.0	-0.9
OpenCQA	6K	5.7	-28.6	-3.6	-2.8	-5.5	-1.7	-7.2	-7.8	-4.1	-1.4	-1.9	-1.6	0.2	-0.5	-12.9	-29.6	-5.4	1.9	-10.4	-8.2	-10.4	-115.6	10.0	-49.3	-47.0	-3.1	7.1	-16.4	10.0	1.5
IconQA	19 K	5.3	-5.1	-1.0	-1.0	-1.8	-0.7	-6.3	-4.5	-4.3	-3.7	-0.9	-0.4	0.9	10.0	-5.6	-29.3	-6.2	10.0	-3.1	-4.2	-1.0	-50.1	-1.6	-14.9	-68.7	-1.9	-6.5	6.1	-159.8	-0.5
GQA	943 K	5.0	-85.3	-30.6	-15.6	-3.0	5.3	-6.7	2.0	-2.9	-2.1	-0.3	10.0	10.0	-0.3	-11.0	-7.3	7.7	1.9	-11.1	-7.6	-1.2	-19.8	-2.3	-13.0	-4.4	-0.1	-9.3	7.7	-230.0	-1.1
OK-VQA	9 K	4.9	-43.6	-15.4	-5.7	-5.7	0.8	10.0	10.0	0.5	0.6	-1.5	-3.0	0.0	-0.4	-8.8	-13.6	-1.0	-0.9	-1.7	-1.7	-0.3	-19.9	-2.2	4.0	-34.4	1.2	-5.9	5.9	-71.5	-0.9
COCO Caption	567 K	4.8	10.0	1.9	1.1	-32.6	-2.1	-16.1	-6.0	-13.9	0.7	-0.4	-9.5	-0.9	1.4	1.5	-9.5	-4.1	0.0	-8.8	-6.7	0.4	-6.7	1.3	-6.0	-21.5	0.4	1.5	-0.7	-59.9	0.8
TextCaps	549 K	4.7	-52.0	-16.5	10.0	-28.0	-3.5	-9.0	-5.7	-5.5	0.7	-0.7	-8.1	-2.6	1.5	-12.4	-9.1	-2.1	0.0	-0.6	4.8	0.0	-0.3	1.3	-2.6	-26.1	0.7	2.9	-9.8	-20.7	0.7
TextVQA	35 K	4.7	-104.8	-39.7	-14.8	-14.7	-8.6	-6.6	-3.5	-6.4	-8.0	-0.7	-6.7	-1.8	-0.7	-31.2	-8.6	-1.2	-3.4	10.0	10.0	0.4	-1.9	-1.8	10.0	-16.9	0.6	-3.2	2.5	-12.0	-0.9
OLIVE	7K	3.9	-94.8	-31.6	-15.5	-18.8	-2.3	-23.1	-5.7	-20.1	-4.4	-0.9	-4.9	-0.3	0.7	-0.2	-23.6	0.0	1.9	-14.4	-1.2	-2.7	-7.2	1.4	-27.0	-10.9	-3.0	10.0	-2.0	-25.7	9.8
HM	9 K	3.3	-10.5	-4.4	-2.1	-33.0	-1.5	-37.9	-4.0	-36.1	-5.6	-1.2	-20.2	-2.5	-0.1	-4.6	-43.0	10.0	-4.3	-23.6	-9.9	-3.3	-85.7	-2.5	-43.3	-66.5	10.0	-10.7	2.0	-231.3	-0.8
VQAv2 QG	444 K	2.9	-80.2	-33.7	-14.8	-27.5	-7.8	-20.4	-7.8	-15.8	-5.4	-2.0	-8.8	-1.4	-0.9	-10.7	-24.6	8.9	4.7	-11.3	-1.7	-3.1	-203.9	1.2	-40.9	-63.9	-2.7	-1.8	-10.0	-71.7	0.8
LLaVA Conversation	57 K	2.4	-56.9	-19.5	-8.4	-59.9	-11.5	-37.7	-23.2	-36.1	-12.8	-3.6	-33.4	-4.2	-0.4	-31.7	-100.1	1.2	1.9	-25.8	-16.8	-16.7	-155.9	7.1	-48.4	-75.2	-3.6	8.9	-29.8	-48.7	10.0
A-OKVQA QG	17 K	2.3	-78.5	-30.0	-11.3	-55.8	-11.8	-26.4	-13.6	-23.3	-5.6	-1.8	-25.9	-2.6	0.4	-8.0	-80.1	5.1	1.9	-17.7	-13.5	-3.7	-198.4	0.2	-35.3	-69.6	0.7	0.1	-29.3	-53.3	0.2
OCR-VQA	802 K	2.3	-12.6	-4.2	-4.7	-29.2	-21.5	-30.1	-20.1	-25.8	-25.4	-2.8	-13.8	-5.5	-1.9	-21.7	-41.6	-6.8	1.9	-16.3	-29.5	10.0	-32.4	-2.5	-38.1	-82.2	-0.9	-11.0	-9.3	-238.2	-1.0
VSR	3 K	2.0	-17.3	-7.0	-1.9	-53.2	-27.3	-35.1	-29.0	-34.4	-17.0	-3.8	-31.3	-4.6	-1.1	-2.2	-112.7	-7.2	1.9	-23.3	-27.8	-11.8	-79.8	-1.7	-50.5	-146.7	-0.4	-10.5	-3.4	-100.5	-1.0
LLaVA Reasoning	77 K	2.0	-98.1	-34.4	-17.3	-67.2	-13.0	-38.5	-21.6	-37.1	-13.2	-1.5	-38.6	-2.3	-1.0	-18.3	-112.3	5.1	0.0	-28.2	-27.7	-18.4	-145.4	-0.8	-51.9	-47.2	-3.6	9.6	-7.5	-48.5	1.1
OK-VQA QG	9 K	1.5	-91.1	-33.1	-15.0	-66.7	-31.3	-37.3	-26.1	-35.7	-24.9	-3.7	-37.8	-12.6	-1.6	-30.2	-113.0	0.7	1.9	-26.5	-24.1	-6.7	-241.5	0.0	-42.3	-85.9	-3.1	-2.6	-22.7	-79.6	0.6
LLaVA Description	23 K	1.3	-95.2	-31.9	-16.4	-67.2	-68.7	-38.5	-58.4	-37.1	-37.5	-3.4	-38.6	-29.3	-3.0	-53.7	-113.2	-27.6	1.9	-28.1	-56.4	-18.4	-298.5	-1.2	-52.8	-155.9	-6.4	8.2	-38.4	-118.0	0.5

Table 8: Normalized transfer learning performance of BLIP-2. Higher values indicate better transferability. The rows are sorted in descending order of average performance. We multiply the values by a factor of 10 to aid visualization. The highest performance in each column is 10. QG denotes question generation, MC denotes multiple-choice and G denotes open-ended generation. The color scale is normalized along each column. The colors represent values in descending order: green, yellow, orange and red.

Source Task	Dataset Size	AHP Ranking Score	Target Task: COCO Caption	Flickr 30k	Text Caps	VQAv2 G	VQAv2 MC	OK-VQA G	OK-VQA MC	A-OKVQA G	A-OKVQA MC	Science QA	GQA G	GQA MC	Icon QA	VSR	CLEVR G	CLEVR MC	RAVEN-FAIR	Text VQA G	Text VQA MC	OCR-VQA G	OCR-VQA MC	Open CQA	Chart QA G	Chart QA MC	HM	NY Explain	NY Rank	MORE	OLIVE
VQAv2	444 K	6.9	3.9	3.1	1.9	10.0	10.0	9.6	9.1	10.0	7.4	3.4	7.9	6.8	3.3	6.8	10.0	6.5	-1.0	8.0	7.9	5.2	1.3	1.8	10.0	-34.9	9.7	-20.3	-4.3	-3.8	-101.0
A-OKVQA (MC)	17 K	6.5	1.3	1.4	1.2	3.2	6.0	3.7	9.1	4.0	10.0	5.3	0.8	5.0	2.6	-15.9	-2.3	2.6	2.4	3.9	10.0	2.0	8.2	3.7	4.9	10.0	3.4	1.3	2.2	0.9	-60.0
ScienceQA	6 K	6.3	0.7	0.5	0.7	3.5	3.7	3.6	6.5	3.8	6.8	10.0	3.0	2.5	2.6	-10.0	1.8	-20.6	6.5	2.9	6.9	2.0	6.9	2.6	6.6	-20.3	-0.3	2.2	10.0	0.6	10.0
GQA	943 K	5.8	2.9	1.7	0.7	6.1	7.8	0.7	7.3	2.4	7.0	2.8	10.0	10.0	2.3	4.3	5.4	10.0	-1.0	-0.2	2.3	3.1	3.0	-1.1	0.9	-24.0	6.9	2.7	5.8	1.0	-100.7
A-OKVQA	17 K	5.8	2.6	1.7	1.1	6.8	8.0	8.4	10.0	9.3	8.0	3.2	5.2	5.8	3.4	-18.0	2.9	9.5	10.0	4.8	7.4	2.7	-1.0	-0.9	6.8	-26.6	4.8	-21.2	0.3	-1.8	-102.1
TextVQA	35 K	5.4	5.7	5.6	4.8	6.1	5.6	4.1	6.7	4.2	6.9	3.8	4.7	3.8	2.4	-2.3	7.9	-8.0	0.1	10.0	8.7	5.3	3.8	0.9	9.9	-27.8	7.2	-12.1	-11.7	-0.4	-97.1
VSR	3 K	5.4	0.0	0.0	0.1	4.3	1.5	1.2	0.7	2.1	0.4	0.2	4.5	1.1	0.7	10.0	3.8	0.1	0.1	1.4	0.4	3.7	3.4	0.5	0.8	4.7	-6.2	-0.4	-0.8	0.2	1.3
Flickr30k	145 K	5.4	7.0	10.0	5.9	1.0	2.5	-0.8	3.3	-0.2	2.9	1.7	2.0	2.3	1.9	0.9	5.6	1.4	-0.9	2.8	5.3	2.1	2.3	5.0	3.4	-14.6	10.0	5.2	0.1	6.0	-35.9
OCR-VQA	802 K	5.0	3.9	3.9	4.1	5.1	0.1	2.3	1.3	3.3	4.2	3.0	5.3	3.8	1.1	-6.7	7.3	-76.7	1.3	5.2	-0.4	10.0	10.0	2.5	7.4	-14.3	7.9	4.6	-7.0	2.4	-94.7
OK-VQA	9 K	5.0	3.3	2.5	1.6	4.5	4.2	10.0	8.6	4.3	5.6	3.4	3.1	4.1	3.1	-18.0	-3.3	-1.6	0.9	3.7	6.5	1.8	2.5	1.0	7.3	-19.9	4.8	-12.8	8.7	3.3	-101.2
IconQA	19 K	4.9	0.5	0.4	0.7	2.8	4.6	3.0	4.0	3.0	4.9	2.9	2.6	3.0	10.0	-9.3	1.8	-14.5	-4.7	2.6	4.4	2.2	7.0	2.9	5.7	-25.8	2.1	3.2	5.7	0.9	-7.9
HM	9 K	4.7	0.0	0.0	0.1	2.2	0.1	-0.6	-0.4	-1.0	-0.1	-1.0	2.0	-0.5	1.0	-11.9	-0.1	-8.1	-0.1	-0.3	-2.6	1.5	2.4	0.2	0.0	-5.3	3.8	-0.3	-6.0	-0.2	8.6
OpenCQA	6K	3.9	1.5	1.1	1.5	0.7	-0.9	-1.2	2.3	-2.0	2.7	2.8	0.0	-0.7	1.8	-9.6	-5.8	-10.4	-0.5	1.1	1.7	0.8	1.8	10.0	0.3	-13.5	-5.2	7.4	-4.3	6.6	-33.8
LLaVA Conversation	57 K	3.8	0.9	0.2	0.8	-8.9	-1.2	-12.0	1.2	-12.8	0.5	1.1	-9.9	-3.2	-0.1	-5.5	-23.7	-39.3	-1.1	-4.5	-0.5	-3.1	0.1	3.2	-3.6	-20.1	8.6	0.3	-12.2	0.0	8.2
COCO Caption	567 K	3.8	10.0	7.9	5.3	-8.6	2.6	-13.8	3.0	-13.4	1.1	-0.3	-9.7	-0.3	1.0	-5.5	-22.7	4.5	-1.1	-5.7	2.9	-2.9	1.8	5.8	-3.5	-33.7	3.4	4.4	-8.7	6.8	-48.4
TextCaps	549 K	3.8	3.8	3.6	10.0	-3.4	-4.6	-6.2	3.0	-4.6	2.4	-0.6	-3.7	-5.5	0.9	-6.8	-3.4	-43.1	-1.6	-0.8	3.9	1.8	1.3	8.3	-0.9	-43.9	-1.7	7.9	-8.3	10.0	-45.1
VQAv2 QG	444 K	3.6	3.9	4.3	1.2	-5.3	1.4	-6.6	0.8	-5.4	3.0	0.3	-6.4	3.2	1.1	-11.7	-16.7	-7.6	-1.0	-2.3	2.5	-1.7	0.6	2.8	-2.5	-23.5	3.1	6.2	-6.6	5.4	-78.1
Web CapFilt	23,147 K	3.6	8.4	8.0	6.3	-5.3	1.2	-7.8	3.7	-5.6	4.8	1.3	-5.9	-0.2	1.9	-11.2	-15.5	-21.0	1.0	-2.8	3.6	-2.7	-0.5	3.0	-4.1	-54.9	-7.9	-1.5	-11.9	7.7	-70.5
LLaVA Reasoning	77 K	3.4	0.1	-0.1	-0.2	1.5	-0.6	-3.5	-0.5	-3.1	-0.6	-0.6	1.7	-4.3	-1.3	-11.3	-4.3	-6.7	-1.3	0.3	-2.0	-3.2	-1.7	-0.3	-4.4	-9.6	2.1	0.8	-6.6	0.4	-96.2
OK-VQA QG	9 K	3.1	1.8	1.4	0.4	-6.2	-2.1	-9.0	-0.9	-7.4	0.9	-1.2	-7.5	-3.7	1.5	-16.8	-13.4	-16.4	-4.6	-3.9	-1.9	-2.4	-0.8	2.7	-3.3	-21.6	-5.5	10.0	-0.5	5.9	-65.3
OLIVE	7K	3.0	-0.1	-0.3	0.1	-1.9	-1.4	-10.2	-1.1	-9.2	0.3	-0.3	-2.5	-0.9	-0.2	-15.8	-13.4	-49.6	-4.1	-2.2	-0.7	-0.9	0.3	2.1	-3.7	-20.5	-5.9	3.5	-1.6	1.0	3.5
A-OKVQA QG	17 K	2.9	0.8	0.2	0.2	-6.4	-1.1	-12.1	-0.4	-9.7	-0.6	-1.6	-7.6	-2.2	0.4	-8.6	-14.4	-8.5	4.2	-4.5	-2.9	-2.5	-1.0	1.2	-3.0	-31.4	-12.8	4.0	-8.8	5.1	-82.7
LLaVA Description	23 K	2.3	-0.1	-0.4	-0.3	-9.7	-4.3	-18.4	-3.7	-16.5	-2.1	-3.9	-10.5	-6.1	-1.5	-12.3	-24.5	-25.5	-7.7	-7.0	-4.9	-3.2	1.4	-0.5	-4.4	-34.2	-5.2	2.2	-6.0	0.0	-100.7

Table 9: Normalized transfer learning performance of LLaVA. Higher values indicate better transferability. The rows are sorted in descending order of average performance. We multiply the values by a factor of 10 to aid visualization. The highest performance in each column is 10. QG denotes question generation, MC denotes multiple-choice and G denotes open-ended generation. The color scale is normalized along each column. The colors represent values in descending order: green, yellow, orange and red.

Source Task	Dataset Size	AHP Ranking Score	Target Task: COCO Caption	Flickr 30k	Text Caps	VQAv2 G	VQAv2 MC	OK-VQA G	OK-VQA MC	A-OKVQA G	A-OKVQA MC	Science QA	GQA G	GQA MC	Icon QA	VSR	CLEVR G	CLEVR MC	RAVEN-FAIR	Text VQA G	Text VQA MC	OCR-VQA G	OCR-VQA MC	Open CQA	Chart QA G	Chart QA MC	HM	NY Explain	NY Rank	MORE	OLIVE
VQAv2	444 K	5.7	1.5	1.1	1.7	10.0	10.0	9.3	8.5	9.5	7.9	2.8	8.9	6.7	0.8	9.8	9.6	1.9	0.7	8.1	7.3	5.9	4.1	-8.5	9.0	-3.9	1.0	-19.5	-5.0	0.5	0.3
A-OKVQA (MC)	17 K	5.6	2.1	1.6	2.6	8.5	5.2	7.7	8.8	7.8	10.0	4.0	7.0	-0.6	1.5	1.3	1.6	-1.2	0.0	5.7	10.0	0.4	3.3	-2.2	5.7	1.6	0.7	6.3	2.5	8.4	3.9
ScienceQA	6 K	5.5	3.8	4.0	3.2	7.1	-1.7	5.9	4.6	5.2	5.2	10.0	6.5	-2.7	1.0	-0.2	6.9	-4.9	10.0	5.3	6.7	0.7	1.4	-8.0	4.5	-11.2	0.6	10.0	10.0	9.6	-0.2
A-OKVQA	17 K	5.3	1.2	0.6	1.2	8.7	5.7	9.4	10.0	10.0	7.7	2.8	7.8	2.0	1.1	9.0	8.8	-1.6	-0.3	7.8	5.7	3.8	3.7	-9.6	8.6	-2.4	1.3	-8.5	-4.2	3.0	0.2
OK-VQA	9 K	5.1	0.4	-0.4	0.9	8.2	3.8	10.0	9.1	8.1	6.6	2.5	7.3	0.1	-0.1	7.6	9.3	-2.3	-5.1	7.6	6.7	2.8	1.3	-9.8	8.2	-1.7	1.5	-14.0	5.8	3.5	0.1
OCR-VQA	802 K	5.0	6.4	5.4	4.1	2.4	4.8	2.4	2.9	2.2	2.8	2.6	1.8	4.2	0.2	8.9	1.6	3.4	0.0	4.9	1.4	10.0	10.0	-5.5	7.8	-0.4	0.2	4.7	-3.3	6.2	1.7
GQA	943 K	4.9	0.2	-0.7	1.0	8.8	7.9	7.0	7.1	7.9	6.6	1.5	10.0	10.0	0.8	0.5	10.0	0.3	0.0	5.7	1.0	4.7	2.0	-9.2	7.7	-4.7	0.8	-6.0	-7.1	2.5	0.2
HM	9 K	4.8	7.3	6.6	4.3	7.0	3.2	3.7	0.7	3.4	1.6	0.7	6.3	2.1	0.2	3.1	6.3	0.2	1.5	3.7	-1.1	4.1	-0.3	-5.8	7.9	-6.1	10.0	-1.6	-9.2	7.9	3.9
OpenCQA	6K	4.6	1.2	0.8	1.1	0.0	1.3	0.0	1.1	0.0	1.2	2.2	0.0	2.0	-0.1	-0.2	0.0	2.1	0.0	0.0	4.0	0.0	1.8	10.0	-2.3	10.0	1.5	3.9	5.0	5.4	5.1
Flickr30k	145 K	4.5	7.8	10.0	6.2	0.1	1.5	0.2	3.2	0.2	3.9	0.6	0.1	1.4	0.3	5.9	0.0	-0.8	5.0	0.8	1.9	0.1	1.9	-2.2	4.1	1.6	0.0	-2.1	-5.6	5.4	2.9
IconQA	19 K	4.5	3.4	2.9	2.9	2.5	-0.8	1.5	0.2	1.5	2.2	0.4	2.1	-0.4	10.0	1.2	5.2	4.2	7.8	1.9	-2.9	1.2	-8.6	-7.9	5.1	-12.3	-0.8	5.8	7.9	7.0	5.0
VSR	3 K	4.5	3.0	2.9	1.2	0.5	2.7	0.4	0.4	0.4	2.7	0.5	0.7	0.1	1.3	10.0	0.0	9.3	0.4	0.2	1.8	0.0	1.5	2.9	1.2	-0.9	0.9	4.5	-5.6	4.6	1.6
TextVQA	35 K	4.5	-0.9	-1.8	0.7	4.9	3.6	6.0	5.0	6.4	4.5	3.2	5.0	3.1	0.9	6.7	8.6	-4.4	1.0	10.0	8.4	6.1	-5.4	-11.0	10.0	-12.3	1.2	-32.0	-2.9	-2.5	0.2
VQAv2 QG	444 K	4.3	5.9	4.9	4.4	1.9	-2.7	2.5	-0.6	2.5	-0.7	0.7	1.3	-2.8	0.8	-0.7	0.7	0.1	0.0	1.9	-2.1	0.2	2.6	-1.3	3.7	-1.1	0.2	4.0	0.2	9.0	0.8
A-OKVQA QG	17 K	4.2	2.2	1.4	2.9	1.1	-1.7	0.1	0.2	0.2	-0.3	0.5	0.9	-2.2	0.6	3.5	0.1	5.1	0.0	0.7	-4.2	0.1	2.8	-2.9	4.8	-2.8	0.8	3.4	1.7	7.5	4.2
OK-VQA QG	9 K	4.2	1.2	0.9	2.2	0.1	-2.5	0.3	-1.2	0.3	0.0	0.4	0.1	-2.3	1.0	-0.9	0.0	4.2	0.0	0.6	0.7	0.0	2.5	0.0	3.5	-2.4	1.9	2.9	1.5	7.1	4.1
COCO Caption	567 K	4.1	10.0	8.5	6.4	0.0	-2.0	0.0	0.2	0.0	-0.2	0.2	0.0	-1.6	0.9	-1.2	0.0	1.8	1.0	-0.1	-1.7	0.1	-0.3	-3.8	-0.2	6.3	0.3	-6.8	-7.1	5.2	2.6
Web CapFilt	23,147 K	4.1	9.3	8.1	6.7	0.2	2.8	0.3	4.3	0.2	4.8	1.1	0.4	3.1	0.0	-2.3	0.0	-0.1	1.0	0.4	2.1	0.2	-7.8	-5.4	-2.3	-15.6	-0.7	-14.3	-5.2	4.6	2.4
OLIVE	7K	4.0	0.1	0.3	0.8	0.0	-2.4	0.0	-0.3	0.0	0.2	0.2	0.0	-1.8	0.4	5.1	0.0	6.7	0.0	0.0	-1.4	0.0	0.5	-0.3	-1.8	-5.4	1.3	-2.4	6.7	2.4	10.0
TextCaps	549 K	4.0	6.4	6.1	10.0	0.1	-0.1	0.2	2.2	0.1	2.3	0.0	0.1	0.8	-0.7	-1.2	0.0	0.2	8.6	0.1	3.0	0.0	-9.4	-2.2	-2.3	-31.1	-0.1	-5.4	-12.1	10.0	2.5
LLaVA Conversation	57 K	3.9	1.4	1.1	1.8	0.0	-13.6	0.0	-1.8	0.0	-1.3	-0.8	0.0	-15.7	0.1	3.5	0.0	10.0	0.0	-0.2	-5.4	0.0	0.8	0.8	-1.7	3.4	2.3	-1.7	-1.0	3.6	-0.1
LLaVA Reasoning	77 K	3.6	-0.1	0.2	0.3	0.0	-4.1	0.0	-0.5	0.0	-0.8	-1.0	0.0	-4.7	-0.2	3.6	0.0	8.4	0.0	-0.1	-5.3	0.0	-1.2	-4.5	-1.9	-4.9	-0.1	-7.2	4.8	0.8	-0.2
LLaVA Description	23 K	3.1	-0.3	-0.1	0.1	0.0	-6.6	0.0	-2.6	0.0	-3.4	-1.8	0.0	-6.6	-0.9	-4.2	0.0	0.7	3.9	-0.2	-11.8	0.0	-3.9	-4.8	-2.3	-7.8	-0.6	-3.3	-4.6	1.7	0.3

Table 10: Normalized transfer learning performance of MiniGPT-4. Higher values indicate better transferability. The rows are sorted in descending order of average performance. We multiply the values by a factor of 10 to aid visualization. The highest performance in each column is 10. QG denotes question generation, MC denotes multiple-choice and G denotes open-ended generation. The color scale is normalized along each column. The colors represent values in descending order: green, yellow, orange and red.

Source Task	Dataset Size	AHP Ranking Score	Target Task																												
			COCO Caption	Flickr 30k	Text Caps	VQAv2		OK-VQA		A-OKVQA		Science QA	GQA		Icon QA	VSR	CLEVR		RAVEN-FAIR	Text VQA		OCR-VQA		Open CQA	Chart QA		HM	NY Explain	NY Rank	MORE	OLIVE
						G	MC	G	MC	G	MC		G	MC			G	MC		G	MC	G	MC		G	MC					
A-OKVQA (MC)	17 K	6.0	6.7	6.4	2.9	8.4	3.1	7.4	9.0	7.7	10.0	5.4	6.5	-2.2	2.9	3.2	5.7	8.0	-0.4	7.1	10.0	5.0	6.3	1.1	8.8	3.4	0.0	6.5	4.3	2.0	2.5
VQAv2	444 K	5.9	4.1	3.6	1.5	10.0	10.0	9.1	9.0	9.7	8.2	4.1	8.8	6.1	2.4	7.5	10.0	8.7	6.1	8.8	9.9	7.0	-1.5	-0.1	9.9	-18.4	2.2	-45.6	2.6	-3.6	-0.6
OCR-VQA	802 K	5.7	7.7	7.9	4.0	7.7	6.8	5.7	4.8	6.8	5.7	3.0	7.4	4.5	0.8	-1.9	9.8	4.6	8.0	6.9	0.4	10.0	10.0	0.5	9.5	1.2	5.4	3.2	5.7	-7.8	-0.6
A-OKVQA	17 K	5.6	3.2	2.8	0.2	9.2	8.9	9.2	10.0	10.0	8.0	4.0	8.4	5.0	3.0	-6.9	9.5	10.0	-0.6	7.9	9.1	6.6	-0.7	-0.2	8.2	-3.2	-0.1	-62.5	10.0	-4.3	-0.7
ScienceQA	6 K	5.5	4.9	5.1	3.3	4.6	5.6	5.1	6.4	4.9	8.7	10.0	5.4	1.5	1.9	3.2	5.3	6.5	7.7	6.4	7.8	5.4	5.7	2.7	5.0	-14.5	-0.1	6.1	-0.4	5.5	0.1
IconQA	19 K	5.2	4.1	4.2	2.4	7.0	5.1	5.3	4.7	5.3	5.6	3.9	6.9	0.1	10.0	2.5	8.8	1.5	9.2	7.3	1.2	6.6	-3.1	3.0	9.8	-23.7	1.8	4.0	-0.1	1.4	1.9
GQA	943 K	5.2	3.5	3.0	-0.3	8.7	8.7	7.0	9.3	7.7	7.6	3.1	10.0	10.0	1.6	-2.5	8.4	6.5	10.0	4.8	4.3	5.4	-3.6	-0.2	6.1	-10.5	-0.1	-53.7	8.0	-6.9	-0.8
VSR	3 K	5.0	3.7	4.1	2.3	5.9	6.3	2.9	4.7	2.5	4.7	1.9	5.8	4.0	1.9	10.0	3.4	2.6	-1.3	5.1	4.5	3.7	8.6	4.8	2.7	-3.5	0.1	0.0	2.6	0.8	2.0
OK-VQA	9 K	4.8	3.5	2.7	0.9	8.6	5.9	10.0	9.4	8.0	6.6	3.7	7.7	3.7	2.0	-6.6	7.9	7.5	-1.6	7.5	7.1	6.4	-8.1	-0.1	9.6	-13.8	-0.1	-58.1	4.8	-4.3	-0.8
TextVQA	35 K	4.7	7.0	7.1	3.4	7.3	4.1	6.3	6.8	6.4	5.8	2.4	6.5	2.7	1.5	1.8	7.9	4.9	-3.5	10.0	5.6	7.0	-9.5	0.2	10.0	-11.3	2.5	-55.5	-9.0	-6.5	-0.6
OpenCQA	6 K	4.6	4.5	4.6	4.1	0.6	1.3	1.4	2.9	0.9	2.1	2.6	0.4	0.7	1.9	-6.1	0.0	2.0	-6.8	3.3	2.6	1.0	2.6	10.0	1.3	-1.3	2.5	10.0	-1.7	8.2	6.4
COCO Caption	567 K	4.6	10.0	8.9	5.2	0.0	0.6	0.0	0.5	0.0	2.2	-0.5	0.0	0.5	0.3	-4.1	0.0	4.9	-0.4	0.1	5.3	0.4	-3.3	4.9	0.1	10.0	-0.1	1.2	-5.7	5.8	6.3
Flickr30k	145 K	4.5	7.5	10.0	5.7	0.5	2.9	1.1	2.1	1.3	3.4	0.1	0.6	1.1	1.6	-5.8	0.0	3.2	3.2	1.3	5.0	1.0	-1.5	5.4	1.4	2.0	0.8	3.8	-5.5	5.3	6.5
TextCaps	549 K	4.5	5.1	5.2	10.0	0.7	-2.1	1.4	1.3	1.4	0.9	-0.1	0.6	-0.3	-1.1	-3.2	0.0	5.2	7.3	2.5	2.8	1.0	-5.6	5.5	0.8	2.1	-0.1	4.2	-5.4	10.0	5.7
HM	9 K	4.1	3.0	2.6	2.5	4.6	2.8	0.5	1.4	0.5	1.6	-0.2	4.3	1.1	2.1	-7.6	5.2	1.6	1.3	2.4	-1.9	5.8	-3.5	5.3	2.2	-9.5	10.0	1.8	-2.8	2.7	5.3
Web CapFilt	23,147 K	4.0	9.3	9.0	6.9	1.7	3.3	2.8	3.3	3.3	3.4	0.1	1.9	3.9	-2.2	-8.2	0.5	1.0	-5.6	2.2	3.4	0.8	-10.6	3.8	1.5	-26.6	0.3	-10.0	-11.6	8.1	2.7
VQAv2 QG	444 K	3.9	7.5	7.3	2.0	0.5	-5.1	0.1	-1.7	0.0	0.4	-1.4	0.3	-3.6	-0.7	-4.7	0.3	7.3	0.0	0.3	-0.9	1.3	-10.1	3.3	0.6	-2.7	-0.1	2.3	1.2	4.1	3.9
A-OKVQA QG	17 K	3.6	5.1	5.1	2.1	0.4	-6.2	0.0	2.0	0.0	0.3	-0.6	0.1	-3.5	-0.6	-4.6	0.0	8.8	-2.9	0.2	-0.5	0.9	-5.3	3.6	0.1	-8.0	1.6	-12.2	-4.9	3.9	3.1
OLIVE	7 K	3.5	-0.2	-0.1	-0.4	0.0	-0.4	0.0	1.3	0.0	0.8	-0.7	0.0	0.6	0.7	0.1	0.0	-9.8	-2.7	0.0	-2.1	0.0	-0.3	5.2	0.0	-5.3	-0.1	-0.6	-7.0	0.8	10.0
LLaVA Conversation	57 K	3.4	3.5	3.2	1.8	0.0	-0.1	0.0	-0.4	0.0	-0.4	-0.6	0.0	-1.6	-0.5	-7.2	0.0	-10.5	0.0	0.0	-1.2	0.0	-0.8	5.0	0.0	-3.1	-0.1	-3.5	-4.1	-0.5	0.2
LLaVA Description	23 K	3.2	-0.7	-0.5	-1.4	0.0	-3.4	0.0	-0.3	0.0	-1.6	-1.6	0.0	-2.3	-1.6	-4.7	0.0	-1.4	-3.2	0.0	-4.9	0.0	-6.9	1.4	0.0	3.1	-0.1	-1.5	-6.8	-0.1	-0.1
OK-VQA QG	9 K	3.2	7.2	6.9	2.6	0.0	-9.6	0.0	1.4	0.0	0.2	-1.3	0.0	-10.0	-0.3	-7.9	0.0	2.0	-0.1	0.1	-2.2	0.2	-12.7	3.9	0.3	-10.6	0.6	-43.0	-6.2	2.9	0.1
LLaVA Reasoning	77 K	3.2	-0.5	-0.2	-1.1	0.0	-2.7	0.0	-2.2	0.0	-1.3	-1.7	0.0	-2.1	-0.9	-0.6	0.0	0.7	0.3	0.0	-5.5	0.0	-3.0	1.5	0.0	-5.1	-0.1	-8.1	-4.6	-1.7	2.7

Table 11: Normalized transfer learning performance of mPLUG-Owl. Higher values indicate better transferability. The rows are sorted in descending order of average performance. We multiply the values by a factor of 10 to aid visualization. The highest performance in each column is 10. QG denotes question generation, MC denotes multiple-choice and G denotes open-ended generation. The color scale is normalized along each column. The colors represent values in descending order: green, yellow, orange and red.

F Mean Cosine Similarity of Target Tasks

In this section, we rank the average cosine similarity among the target tasks. We first compute pairwise cosine similarity using the SVD features of target tasks. For each target task, we take the average of all pairs that it is involved in. Finally, we rank all target tasks in a descending order.

Target Task	Cosine Similarity
OK-VQA (MC)	0.54
VQAv2 (G)	0.54
A-OKVQA (MC)	0.54
GQA (G)	0.53
OK-VQA (G)	0.53
A-OKVQA (G)	0.53
TextVQA (MC)	0.52
VQAv2 (MC)	0.51
TextVQA (G)	0.50
CLEVR (G)	0.50
ChartQA (G)	0.48
OCR-VQA (G)	0.45
ScienceQA (MC)	0.45
GQA (MC)	0.44
Hateful Memes (MC)	0.42
VSR (MC)	0.41
NY Ranking (MC)	0.37
IconQA (MC)	0.35
OCR-VQA (MC)	0.32
TextCaps (G)	0.29
RAVEN-FAIR (MC)	0.27
Flickr30k (G)	0.25
COCO Caption (G)	0.22
ChartQA (MC)	0.15
MORE (G)	0.15
CLEVR (MC)	0.14
OLIVE	-0.06
NY Explanation (G)	-0.09
OpenCQA (G)	-0.27

Table 12: Mean cosine similarity, computed from the SVD features, for each target task. The tasks are ranking by descending similarity.

G Hierarchical Clustering of SVD Similarity

In this section, we perform hierarchical clustering on the SVD similarity features of target tasks using the Ward's linkage criterion which minimizes the total intra-cluster variance. In Figure 5, we show that hierarchical clustering forms meaningful clusters. For example, captioning tasks are clustered together. Generative and multiple-choice evaluated target tasks are grouped into different groups. This cluster supports the generative vs multiple-choice evaluation factor from factor analysis. However, the clusters are not as comprehensive as common factors extracted by factor analysis. For example, hierarchical clustering does not elucidate factors such as reading vs reasoning, and spatial reasoning.

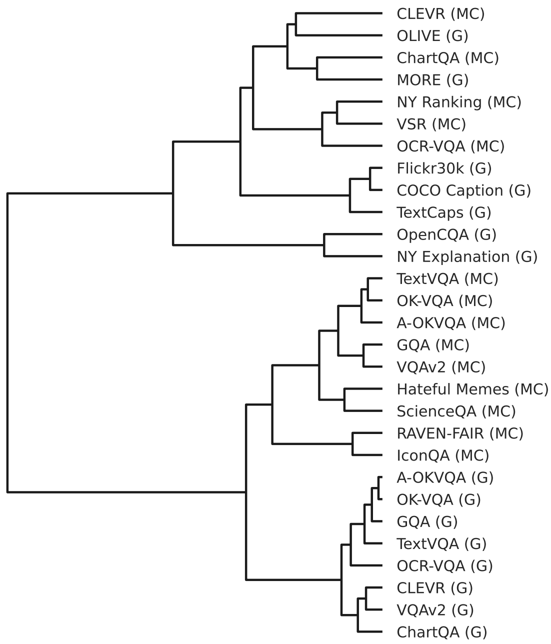


Figure 5: Hierarchical clustering of target tasks.

H Factor Analysis Details

Here we show all the factor loadings of the six factors from the residual matrix $\bar{A}$. Communality quantifies the proportion of variance in each target task that is accounted for by the identified factors. A low communality value indicates that a task differs significantly from others in the mix.

Target Tasks	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Communality
Flickr30k	0.97	-0.02	0.00	0.07	0.06	-0.08	0.96
COCO Caption	0.93	-0.05	0.00	0.10	-0.02	-0.12	0.90
TextCaps	0.83	0.12	-0.20	0.07	0.10	-0.10	0.77
TextVQA (G)	-0.19	0.87	0.04	-0.10	-0.14	-0.16	0.85
VQAv2 (MC)	-0.34	-0.74	-0.34	-0.01	0.24	-0.02	0.83
ChartQA (G)	-0.08	0.67	-0.16	0.31	-0.12	-0.23	0.65
OK-VQA (G)	-0.24	0.60	0.51	-0.20	0.20	0.15	0.78
GQA (MC)	-0.32	-0.55	-0.18	-0.26	-0.02	0.00	0.50
OK-VQA (MC)	-0.43	-0.49	-0.30	0.07	0.22	-0.20	0.62
VQAv2 (G)	0.08	0.06	0.85	0.23	0.05	-0.25	0.86
GQA (G)	-0.22	-0.01	0.75	-0.05	-0.21	0.12	0.66
A-OKVQA (G)	-0.28	0.54	0.59	-0.26	0.23	0.17	0.87
TextVQA (MC)	-0.38	-0.12	-0.49	0.02	0.36	-0.23	0.58
OCR-VQA (MC)	0.20	-0.14	-0.04	0.65	-0.19	-0.27	0.60
ChartQA (MC)	-0.14	0.07	-0.02	0.65	0.19	0.29	0.57
RAVEN-FAIR (MC)	0.02	-0.01	0.08	-0.40	-0.04	0.17	0.20
ScienceQA (MC)	-0.07	0.00	-0.07	-0.39	-0.05	-0.06	0.17
IconQA (MC)	-0.01	-0.09	-0.08	-0.34	-0.05	-0.10	0.14
OCR-VQA (G)	-0.01	0.11	-0.04	-0.12	-0.66	0.01	0.46
A-OKVQA (MC)	-0.21	-0.35	-0.38	-0.18	0.63	-0.07	0.74
MORE (G)	0.22	0.47	-0.22	0.21	0.54	-0.03	0.65
OpenCQA (G)	0.17	-0.07	-0.09	0.11	0.32	-0.24	0.21
OLIVE (G)	-0.05	0.06	0.09	0.10	-0.08	0.61	0.40
CLEVR (G)	-0.17	0.20	0.16	-0.44	-0.34	0.59	0.74
CLEVR (MC)	-0.18	-0.13	-0.05	-0.07	0.01	0.55	0.36
VSR (MC)	0.15	-0.26	-0.10	0.10	-0.06	0.50	0.37
NY Explanation (G)	0.13	-0.03	-0.04	0.26	0.21	-0.10	0.14
NY Ranking (MC)	-0.24	-0.30	0.13	0.08	-0.23	0.04	0.22
Hateful Memes (MC)	0.05	-0.09	-0.16	-0.14	-0.24	0.05	0.12

Table 13: Results of EFA on the residuals $\bar{A}$. Cut-off for factor loadings = 0.3.

Target Tasks	Factor 1	Factor 2	Factor 3	Communality
OK-VQA (G)	0.78	0.43	0.44	1.00
A-OKVQA (G)	0.74	0.44	0.49	0.98
ChartQA (G)	0.59	0.68	0.31	0.91
TextVQA (G)	0.63	0.66	0.38	0.97
OCR-VQA (G)	0.30	0.65	0.46	0.73
GQA (G)	0.51	0.46	0.73	1.00
VQAv2 (G)	0.60	0.46	0.60	0.93

Table 14: Results of EFA on generative VQAs. Cut-off for factor loadings = 0.6.

Target Tasks	Factor 1	Factor 2	Factor 3	Communality
OCR-VQA (MC)	0.81	0.31	0.28	0.82
ChartQA (MC)	0.72	0.38	0.21	0.70
A-OKVQA (MC)	0.51	0.69	0.44	0.93
TextVQA (MC)	0.53	0.69	0.39	0.90
OK-VQA (MC)	0.59	0.64	0.44	0.95
GQA (MC)	0.23	0.28	0.93	1.00
VQAv2 (MC)	0.50	0.55	0.64	0.96

Table 15: Results of EFA on multiple-choice VQAS $\bar{A}$. Cut-off for factor loadings = 0.6.