TOWARD DATA-DRIVEN SKILL IDENTIFICATION FOR GENERAL-PURPOSE VISION-LANGUAGE MODELS

Anthony Meng Huat Tiong[1*], **Junqi Zhao**[1*], **Boyang Li**[1♠],
Junnan Li, Steven C.H. Hoi[2], and **Caiming Xiong**[3]
[1]Nanyang Technological University [2]Singapore Management University [3]Salesforce Research
{anthonym001, junqi.zhao, boyang.li}@ntu.edu.sg

ABSTRACT

The evolution of vision-language (VL) models towards broad competencies has complicated benchmarking, necessitating diverse tasks for accurate evaluation. Moving beyond intuition-guided task selection common in existing benchmarks, we propose a data-driven approach that leverages transfer performance and Factor Analysis (FA) to identify latent skills crucial for VL tasks. Our study demonstrates the utility of FA in systematically understanding and evaluating VL models.

1 INTRODUCTION

Recently developed vision-language models (VLMs) (Dai et al., 2023; Zhu et al., 2023; Liu et al., 2023c; Ye et al., 2023; Li et al., 2023a; Awadalla et al., 2023), viewed as precursors to general-intelligence systems, demonstrate competencies across diverse VL tasks. These tasks are believed to be underpinned by atomic VL skills such as object recognition, spatial relationship recognition, language grounding, etc. However, these skills are not directly observable, and often, a single skill may support several disparate tasks. Current benchmarks (Bitton et al., 2023; Xu et al., 2023; Liu et al., 2023d; Yu et al., 2023; Li et al., 2023c; Fu et al., 2023; Bai et al., 2023b) rely heavily on the intuition of their designers to select test tasks and categorize them by skill, leading to uncertainties about whether these benchmarks fully capture all relevant skills or weight them equally in evaluations.

In this paper, we propose a data-driven approach to identify latent VL skills from performance data. We finetune four VLMs—BLIP-2 (Li et al., 2023b), Mini-GPT4 (Zhu et al., 2023), LLaVA (Liu et al., 2023c), and mPLUG-Owl (Ye et al., 2023)—across 23 source tasks and evaluate them on 29 target tasks, resulting in 2,784 performance measurements, inclusive of 29×4 zero-shot performances. Through Factor Analysis (FA) (Spearman, 1904), a widely used statistical method for uncovering latent factors in data, we identify six VL skills that, upon close inspection, reveal prominent data patterns. Our study highlights the capability of FA in identifying unexpected yet coherent skills that influence VLM behavior, paving the way for more systematically designed benchmarks.

2 ANALYSIS TECHNIQUE

We collect data for FA by creating a diverse set of models and observing their performance on a broad range of test tasks in a transfer learning setting. Specifically, we finetune M pretrained models on N source tasks, yielding MN models, and evaluate them on K target tasks. Each source task is assumed to impart a specific set of VL skills to a model, thereby enhancing its performance on target tasks that require those same skills.

The normalized transfer performance of each model m is stored in matrix $A^{(m)} \in \mathbb{R}^{N\times K}$, where each row represents a source task i and each column represents a target task j. To calculate the entries $a_{i,j}^{(m)}$ of $A^{(m)}$, we use the following formula:

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

*Equal contributions. ♠Corresponding Author.

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

Intuitive Category	Task	Source	Target
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: List of source and target tasks used in experiments. G and MC indicate the generative and multiple-choice versions of VQA tasks respectively.

where $b_{i,j}^{(m)}$ is the transfer performance from source task i to target task j, and $b_{0,j}^{(m)}$ is the zero-shot performance on target task j. The best source task, typically the in-domain training task, achieves $a_{i,j}^{(m)} = 1$. After normalization, we concatenate the M matrices along the source-task (row) dimension to form performance matrix A.

2.1 Factor Analysis

Given that tasks requiring the same VL skills are expected to show correlated performances, we apply Exploratory Factor Analysis (EFA) to uncover these latent skills. Mathematically, we treat the jth column of A, $\boldsymbol{a}_j \in \mathbb{R}^{4N}$, as the characteristics of target task j and explain it using L latent factors,

$$\boldsymbol{a}_j = W\boldsymbol{h}_j + \boldsymbol{\mu} + \boldsymbol{\epsilon}, \tag{2}$$

where $W \in \mathbb{R}^{4N \times L}$ reflects how source tasks load onto the L factors and $\boldsymbol{h}_j \in \mathbb{R}^L$ reflects how the target task j decomposes into these factors. $\boldsymbol{\mu}$ is the average vector across target tasks and $\boldsymbol{\epsilon}$ is Gaussian noise. EFA is closely related to PCA, differing mainly in that it assumes a diagonal covariance matrix for $\boldsymbol{\epsilon}$, while PCA assumes a spherical covariance. We apply Varimax rotation (Kaiser, 1958) to concentrate $\boldsymbol{h}_j$ on as few factors as possible for easy interpretation.

Our initial analysis finds that captioning and most VQA tasks predominantly load onto a single factor, likely indicative of a general VL skill. To explore more specific VL competencies, we isolate this dominant factor using a one-factor EFA, transforming the matrix W into a 4N-dimensional vector w. Then, we perform linear regression from $\boldsymbol{w}$ to A by solving the equation:

$$\text{minimize} \quad \left\| A - \boldsymbol{w}\boldsymbol{\beta}^\top - \boldsymbol{\gamma}\mathbf{1}^\top \right\|_F^2, \tag{3}$$

where $\boldsymbol{\beta}, \boldsymbol{\gamma} \in \mathbb{R}^{4N}$ are trainable parameters.

Finally, we conduct EFA on the residuals, $\bar{A} = A - \boldsymbol{w}\boldsymbol{\beta}^\top - \boldsymbol{\gamma}\mathbf{1}^\top$, which contain information about fine-grained VL skills beyond the general VL skill. Using parallel analysis and Velicer's Minimum Average Partial test, we determine the optimal number of factors to be six.

3 Source and Target Tasks

We gather 27 publicly available VL datasets and create variations, yielding 23 source tasks and 29 target tasks, outlined in Table 1. These tasks span categories such as image captioning, various forms

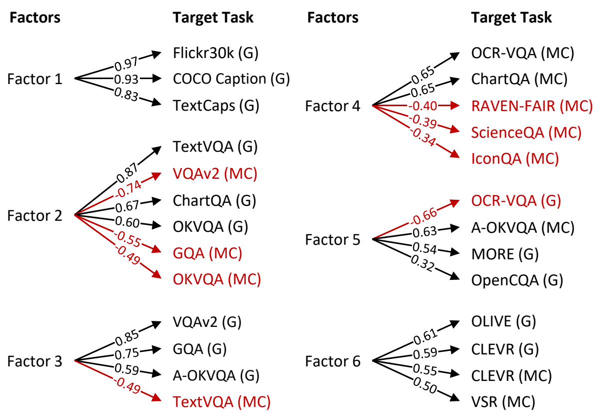

Figure 1: Results of EFA on the residuals $\bar{A}$. Black arrows indicate positive loadings; red arrows indicate negative loadings. Cut-off for factor loadings=0.3.

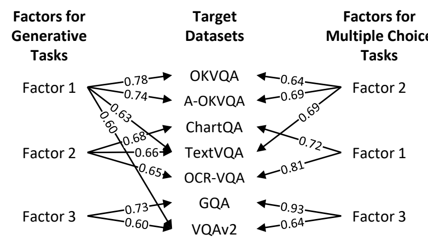

Figure 2: Results of EFA when 3 factors are extracted from 7 generative and 7 MC VQA tasks separately. We merge the results for display. Cut-off for factor loadings=0.6.

of visual question-answering (VQA), visual reasoning, image classification, question and open-ended generation, and humor and sarcasm understanding. Multiple-choice (MC) versions are also created for generative VQA datasets. For detailed descriptions of tasks, see Appendix A.1.

Additionally, we introduce Open-world Language Instruction for Visual-language Evaluation (OLIVE), a new dataset designed to mirror complex user queries encountered by VLMs in real-world scenarios. This type of data is currently underrepresented in existing benchmarks. Generated using ChatGPT and refined through human curation, OLIVE features 9,450 images, 30,120 unique instructions, and 47,250 responses. Further details are in Appendices A.2 and A.3.

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 separate optical character recognition.

4 Experimental Setup and Results

We consider four popular VLMs with limited exposure to our datasets of interest: BLIP-2 (Li et al., 2023b), MiniGPT-4 (Zhu et al., 2023), LLaVA (Liu et al., 2023c), and mPLUG-Owl (Ye et al., 2023). During finetuning, we adjust only the trainable parameters from the VL pretraining phase of each model. The models are trained on each source task for 10,000 iterations, with batch sizes set at 192 for BLIP-2 and 128 for MiniGPT-4, mPLUG-Owl, and LLaVA. Additional details can be found in Appendices A.4 and A.5.

We run EFA on the residual matrix $\bar{A}$ and plot the most significant factor loading for each target task in Fig. 1. A higher absolute factor loading indicates a stronger relationship between the task and the factor. We set the cut-off at 0.3. Notably, New Yorker Explanation and Ranking, and Hateful Memes have no factor loadings above 0.3, implying that they do not share with other tasks any VL skills that are detectable by EFA. Detailed results are provided in Appendices A.6 and A.7.

Furthermore, we analyze the effect of output length on transfer performance by dividing the tasks into three groups based on average output lengths: 1-3 words, 6-12 words, and over 40 words. The average normalized transfer performance by group is shown in Tab. 2.

5 Discussion

In this section, we provide interpretations for the factors identified in factor analysis and highlight key findings from our experiments.

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**

Table 2: Mean normalized transfer performance across tasks, grouped by mean output length. Left values consider all source tasks in a group; right values consider only the top 5 source tasks in a group. In-domain source tasks are excluded.

Factor 1: Captioning. Factor 1 separates the captioning tasks — COCO Caption, Flickr30k, and TextCaps — from the rest of the target tasks, which are mostly VQA tasks.

Factors 2 & 3: Generative vs. MC evaluation. Factors 2 and 3 distinguish between generative and multiple-choice (MC) evaluations for VQA tasks. Generative VQAs exhibit positive loadings on both factors while MC VQAs exhibit negative loadings. Furthermore, specialized VQAs requiring OCR skills, such as TextVQA and ChartQA, load positively on Factor 2 and negatively on Factor 3, whereas generic VQAs, such as VQAv2 and GQA, show the opposite pattern.

These differences arise because generative evaluations require exact matches to ground-truth answers, risking false negatives for valid answers phrased differently, unlike MC evaluations which compare average word probabilities, eliminating the need for strict matches. The exact match requirement also makes generative evaluations more sensitive to output lengths.

Interestingly, separate analyses of generative and MC VQA tasks reveal remarkably similar structures (Fig. 2), identifying factors related to knowledge-based VQA, OCR, and generic or spatial relations. This highlights the efficacy of EFA in capturing underlying structures with appropriate data.

Factors 4 & 5: Text Reading vs Reasoning. Factors 4 and 5 differentiate between tasks involving simple text extraction from those requiring complex multi-hop reasoning. OCR-VQA and ChartQA, which primarily focus on reading text and numbers from images, exhibit loadings opposite to RAVEN-FAIR, ScienceQA, and IconQA, which demand strong logical reasoning skills, and opposite to A-OKVQA, MORE, and OpenCQA, which necessitate the use of external knowledge and contextual understanding. The fact that EFA can find these reasonable skills illustrates 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. This suggests that while OLIVE does require spatial reasoning skills, they account for only a small portion of the overall skills needed. The remaining variance in OLIVE is not explained by the factors identified in our analysis.

Humor, sarcasm, and abstract reasoning remain difficult. The models we tested face difficulties in comprehending humor and sarcasm, as evidenced by their performance on the New Yorker datasets and MORE. In addition, they 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.

The output length bias. Tab. 2 reveals a significant correlation between output length and transfer performance. A mismatch in output lengths between source and target tasks leads to much worse performance. This surprising finding shows that output length could be a shortcut feature for VLMs, suggesting that future benchmarks may need a balance of tasks with varying output lengths. Singhal et al. (2023) find a similar trend in the context of Reinforcement Learning from Human Feedback.

Recommendations. Based on the analysis, we make the following recommendations for future VL benchmarks: (1) both generative and MC assessment should be used; (2) future benchmarks should assign higher priority to tasks that require abstract and deep reasoning; (3) model rankings based on latent skills rather than simple mean or intuitive categorization may be informative.

6 Acknowledgments

Anthony Meng Huat Tiong is supported by Salesforce and Singapore Economic Development Board under the Industrial Postgraduate Programme. Junqi Zhao and Boyang Li are supported by the Nanyang Associate Professorship and the National Research Foundation Fellowship (NRF-NRFF13-2021-0006), Singapore.

References

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

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

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

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

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

Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. 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)*, pp. 18030–18040, June 2022. URL `https://openaccess.thecvf.com/content/CVPR2022/html/Chen_VisualGPT_Data-Efficient_Adaptation_of_Pretrained_Language_Models_for_Image_Captioning_CVPR_2022_paper.html`.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL `https://lmsys.org/blog/2023-03-30-vicuna/`.

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

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

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

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

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. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv 2306.13394*, 2023.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 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)*, pp. 6904–6913, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Goyal_Making_the_v_CVPR_2017_paper.html.

Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 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)*, pp. 688–714, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.41. URL https://aclanthology.org/2023.acl-long.41.

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

Drew A. Hudson and Christopher D. Manning. 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*, pp. 6700–6709. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00686. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Hudson_GQA_A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_CVPR_2019_paper.html.

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

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

Shankar Kantharaj, Xuan Long Do, Rixie Tiffany Leong, Jia Qing Tan, Enamul Hoque, and Shafiq Joty. OpenCQA: Open-ended question answering with charts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11817–11837, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.811. URL https://aclanthology.org/2022.emnlp-main.811.

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

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

Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C. H. Hoi. Lavis: A library for language-vision intelligence, 2022a. URL https://arxiv.org/abs/2209.09019.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022b. URL https://arxiv.org/abs/2201.12086.

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

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. Reform-eval: Evaluating large vision language models via unified re-formulation of task-oriented benchmarks. *arXiv 2310.02569*, 2023c.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48. URL `https://doi.org/10.1007/978-3-319-10602-1_48`.

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

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

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

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

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

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

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: a visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3195–3204, 2019. doi: 10.1109/CVPR.2019.00331. URL `http://openaccess.thecvf.com/content_CVPR_2019/html/Marino_OK-VQA_A_Visual_Question_Answering_Benchmark_Requiring_External_Knowledge_CVPR_2019_paper.html`.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.177. URL `https://aclanthology.org/2022.findings-acl.177`.

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

OpenAI. Chatgpt. 2023a. URL `https://openai.com/chatgpt`.

OpenAI. Gpt-4. 2023b. URL `https://openai.com/research/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. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/radford21a.html`.

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

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

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *ECCV*, 2022.

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

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in rlhf. *arXiv*, 2023.

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

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

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

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

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

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

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

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

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

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

A Appendix

A.1 Dataset Details

We considered 27 publicly available VL datasets and created variations, resulting in 23 source tasks and 29 target tasks. The descriptions of these tasks are provided below.

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). We consider two classic datasets: COCO Caption (Lin et al., 2014) and Flickr30k (Young et al., 2014). Additionally, we incorporate TextCaps, which focuses on describing text within images, and Web CapFilt——a dataset of synthetic captions across diverse 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 highly popular image-text task favored for its versatile question-answering format. VQAv2 (Goyal et al., 2017) is a major benchmark, with more than 200,000 COCO images and 1 million questions. Variants include knowledge-grounded VQA, OCR VQA, Chart VQA, among others, which we'll delve into subsequently.

Measuring performance in VQA can be challenging due to the presence of multiple correct answers for the same question. To address this, we establish two target tasks for each VQA dataset. The first task, the generative (G) version, demands an exact match with one of the ground-truth answers to be considered correct. The second task, the multiple-choice (MC) version, involves the model selecting one option from five choices.

Converting a generative VQA dataset to the MC version involves creating five options for each question, including up to two correct answers to accommodate linguistic variations. We then incorporate incorrect choices by sampling answers from other questions, selecting those with top-k probabilities based on InstructBLIP (Dai et al., 2023). During inference, all options are fed to the model, and the one with the highest average word probability is chosen as the model's 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) centers on science textbook content. OK-VQA (Marino et al., 2019) emphasizes visual recognition and knowledge recall, whereas A-OKVQA (Schwenk et al., 2022) usually necessitates an additional step of reasoning.

OCR VQA. TextVQA (Singh et al., 2019) and OCR-VQA (Mishra et al., 2019) are two VQA datasets that entail recognizing text within images. OCR-VQA focuses on reading text from book covers, whereas TextVQA 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 focuses on data extraction and comparison using short answers.

Visual Reasoning. The term "reasoning" in the VLM literature encompasses various tasks, ranging from simple ones like counting (*e.g., how many apples are in the image?*) to more complex ones involving spatial relations and grounding (*e.g., what is adjacent to the cylinder?*), and may even extend to logical or algebraic operations. In this category, we include five datasets, GQA (Hudson & Manning, 2019), VSR (Liu et al., 2023a), CLEVR (Johnson et al., 2017), IconQA (Lu et al., 2021), and Raven-Fair (Benny et al., 2021). GQA and VSR predominantly feature natural images, while IconQA presents cartoons, and Raven-Fair involves abstract diagrams. CLEVR comprises synthetically rendered images of 3D objects. Among these, Raven-Fair poses the greatest challenge, as its data are derived from the Raven's test (Raven, 1938), an intelligence test originally devised for humans that demands intricate reasoning skills.

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. According to (Hessel et al., 2023), VLMs exhibit poor performance in humor understanding. From their paper, we adopt the New Yorker Ranking task, which involves selecting the best humorous caption for a cartoon from the New Yorker magazine, and the New Yorker Explanation task, which requires the model to explain why the cartoon and its caption evoke humor. Additionally, 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, and A-OKVQA, for the task of question generation from an image and an answer. Additionally, we utilize three subsets of LLaVA-Instruct-150K (Liu et al., 2023c), which respectively emphasize free-form conversation, detailed description, and reasoning. While we employ these as source tasks, we refrain from considering them as target tasks, as these artificial datasets may not accurately represent real-world use cases.

Figure 3: Examples of the OLIVE benchmark for different categories. From left to right: visual recognition, knowledge-based, and creative writing.

A.2 OLIVE

We introduce a new dataset, Open-world Language Instruction for Visual-language Evaluation (OLIVE), designed to simulate the complex user queries that VLMs encounter in real-world scenarios. This highly diverse, human-corrected multimodal dataset aims to mirror the variety and idiosyncrasies of inputs that VLMs receive once deployed publicly, which is currently underrepresented in academic datasets. OLIVE enables the training and evaluation of VLMs in conditions that more closely resemble their ultimate use case.

OLIVE comprises 9,450 images, 30,120 unique instructions, and 47,250 responses. The examples can be broadly categorized into 4 groups: visual recognition, creative writing, knowledge-based, and elaborated description. Tab. 3 shows some examples.

The data curation process begins with random sampling of 9,450 images from LAION-Aesthetics (Schuhmann et al., 2022). We combine the original LAION captions - which may contain entity-specific knowledge - with additional captions generated by BLIP-2 (Li et al., 2023b) to form comprehensive image descriptions. These descriptions are then used as input for the text-only version of ChatGPT to generate tailored instructions and five corresponding responses per instruction. We also manually annotate a few seed examples for each aforementioned categories, and use these as in-context examples to guide ChatGPT. Examples of input prompts are given in Appendix A.3.

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 hired human annotators through Flitto, an annotation company, to thoroughly inspect and correct erroneous data. They are task to: 1) ensure that the instructions contain minimal shortcut information, which would 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.

A.3 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 that 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 '###'.

A.4 Model Details

We experiment with the following four VLMs:

- 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 utilizes 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 between 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.
- LLaVA utilizes 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 linear layer between the image encoder and the LLM. All parameters are frozen except the linear layer and LoRA (Hu et al., 2022) parameters in the LLM. The total and trainable parameters are 7B and 164M respectively.
- mPLUG-Owl utilizes 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 in the LLM. The total and trainable parameters are 7B and 4M respectively.

The four models have largely not been trained on our datasets of interest, with a few exceptions. Specifically, BLIP-2 and MiniGPT-4 have been exposed to COCO Caption and Web CapFilt. mPLUG-Owl has been exposed to COCO Caption. LLaVA has been exposed to its own three 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).

A.5 Training Hyperparameters

We fine-tune the models on each source task using instruction-formatted inputs. Only a single instruction template is used for each task, as preliminary findings indicate that multiple templates degrade performance.

We keep the hyperparameters constant for experiments using the same model architecture. The number of training iterations is set to be 10,000, and the batch sizes are set to be 192 for BLIP-2 and 128 for MiniGPT-4, mPLUG-Owl, and LLaVA.

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 $1e{-}8$ to $1e{-}5$ 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 increased from 0 to $2e{-}5$ in the first 200 steps and then cosine decayed 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 requires 2 hours for training and 2 hours for evaluation. We use the LAVIS (Li et al., 2022a) library for training BLIP-2, MiniGPT-4, and mPLUG-Owl and the original author's codebase for training LLaVA. All evaluations are performed using LAVIS.

A.6 Results: Transfer Performance

We present the experimental results for all four models. Tables 3-6 contains the raw transfer performance. Each row represents a source task and each column represents a target task. Tables 7-10 contains the normalized transfer performance. The rows (source tasks) are sorted in descending order of average performance.

Source Task	Dataset Size	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					
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 3: 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 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	-	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 4: 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	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					
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 G	VQAv2 MC	OK-VQA G	OK-VQA MC	A-OKVQA G	A-OKVQA MC	Science QA	GQA G	GQA MC	Icon QA	VSR
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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

Source Task	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	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	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	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	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	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	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	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	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)	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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 6: 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 7: 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		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					
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 8: 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					
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		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 10: 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.

Among the source tasks, LLaVA Conversation demonstrates significant transfer to OLIVE for both BLIP-2 and LLaVA models. This notable transferability likely stems from the similarities in data distribution between LLaVA Conversation and OLIVE, as both utilize instruction-response pairs generated by OpenAI GPT models (OpenAI, 2023a;b). The key difference is that OLIVE is inspected by human annotators to rectify erroneous data, whereas LLaVA Conversation is not.

A.7 Results: Factor Analysis

We present the results for six-factor EFA on the residual matrix $\bar{A}$. We set the cut-off for factor loadings to be 0.3. 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 11: Results of EFA on the residuals $\bar{A}$. Cut-off for factor loadings = 0.3.

We present the results for three-factor EFA on the normalized transfer performance for the generative and MC versions of seven VQA tasks. We set the cut-off for factor loadings to be 0.6.

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 12: 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 13: Results of EFA on multiple-choice VQAS $\bar{A}$. Cut-off for factor loadings = 0.6.