BLIP3-KALE: Knowledge Augmented Large-Scale Dense Captions

Anas Awadalla[1,2*] **Le Xue**[2] **Manli Shu**[2] **An Yan**[2] **Jun Wang**[2] **Senthil Purushwalkam**[2]
Sheng Shen[4] **Hannah Lee**[1] **Oscar Lo**[1] **Jae Sung Park**[1] **Etash Guha**[1]
Silvio Savarese[◇,2,3] **Ludwig Schmidt**[◇,1] **Yejin Choi**[◇,1] **Caiming Xiong**[◇,2] **Ran Xu**[◇,2]

[1] University of Washington, [2] Salesforce Research, [3] Stanford University,
[4] University of California, Berkeley, [◇]Senior Authors

Abstract

We introduce BLIP3-KALE, a dataset of 218 million image-text pairs that bridges the gap between descriptive synthetic captions and factual web-scale alt-text. KALE augments synthetic dense image captions with web-scale alt-text to generate factually grounded image captions. Our two-stage approach leverages large vision-language models and language models to create knowledge-augmented captions, which are then used to train a specialized VLM for scaling up the dataset. We train vision-language models on KALE and demonstrate improvements on vision-language tasks. Our experiments show the utility of KALE for training more capable and knowledgeable multimodal models.

1. Introduction

We introduce BLIP3-KALE, a dataset of 218 million image-text pairs that advances the knowledge-augmented image captioning. KALE builds upon recent work in this area, particularly CapsFusion [29], which pioneered the use of large language models to fuse synthetically generated image captions with alt-text to incorporate real-world knowledge. KALE makes two key contributions beyond CapsFusion:

Scale and Density: CapsFusion produced 120M samples with an average of 22.74 words per caption. KALE is significantly larger and denser, containing 218M samples with an average of 67.26 words per caption - 1.82x the scale and nearly 3x the caption density.

Efficient Generation: Instead of utilizing multiple large models to synthesize image-captions and fuse the knowledge, we distill the knowledge augmentation process into a compact 2B parameter alt-text-conditioned captioning model. This enables efficient generation of high-quality captions, which allows us to scale up the dataset creation process significantly.

Our approach combines synthetic captions from VLMs with factual information from web-scale alt-text, creating rich image descriptions. We demonstrate that training on KALE improves performance across multimodal tasks compared to many previous purely synthetic or web-scraped datasets.

2. Approach

2.1. Stage 1: Generating initial knowledge-augmented captions

The first stage of our approach focuses on creating an initial pool of dense knowledge-augmented captions. We begin by leveraging CogVLM-17B [26] to generate dense captions for images from the Datacomp-1B dataset. These captions serve as a foundation for our knowledge augmentation process. To enhance these captions with real-world knowledge, we employ Mistral[8], a powerful language model. We prompt Mistral using the CogVLM-generated captions, instructing it to augment the descriptions with relevant factual information. This step aims to incorporate broader contextual knowledge into the image descriptions, following CapsFusions' prompting method. Through this process, we create an initial pool of 100 million knowledge-augmented captions.

2.2. Stage 2: Scaling up

The second stage of our approach focuses on scaling up the dataset to our target of 218 million image-text pairs. We accomplish this by training a specialized VLM using the knowledge-augmented captions generated in Stage 1. We construct our VLM similar to the LLaVA [17] model, using Qwen1.5-1.5B [28] for the language model and DFN ViT-H [6] for the vision encoder. We resize the positional embeddings for the vision encoder to handle 490x490 images matching the resolution used by CogVLM. Our VLM takes two inputs: image patch embeddings and the original Datacomp-1B captions. The model is trained to output the knowledge-augmented captions produced in Stage 1. We use this VLM to caption an additional 118 million images from the Datacomp-1B dataset. This two-stage approach enables us to efficiently scale up KALE to 218 million samples.

*Work done while interning at Salesforce Research

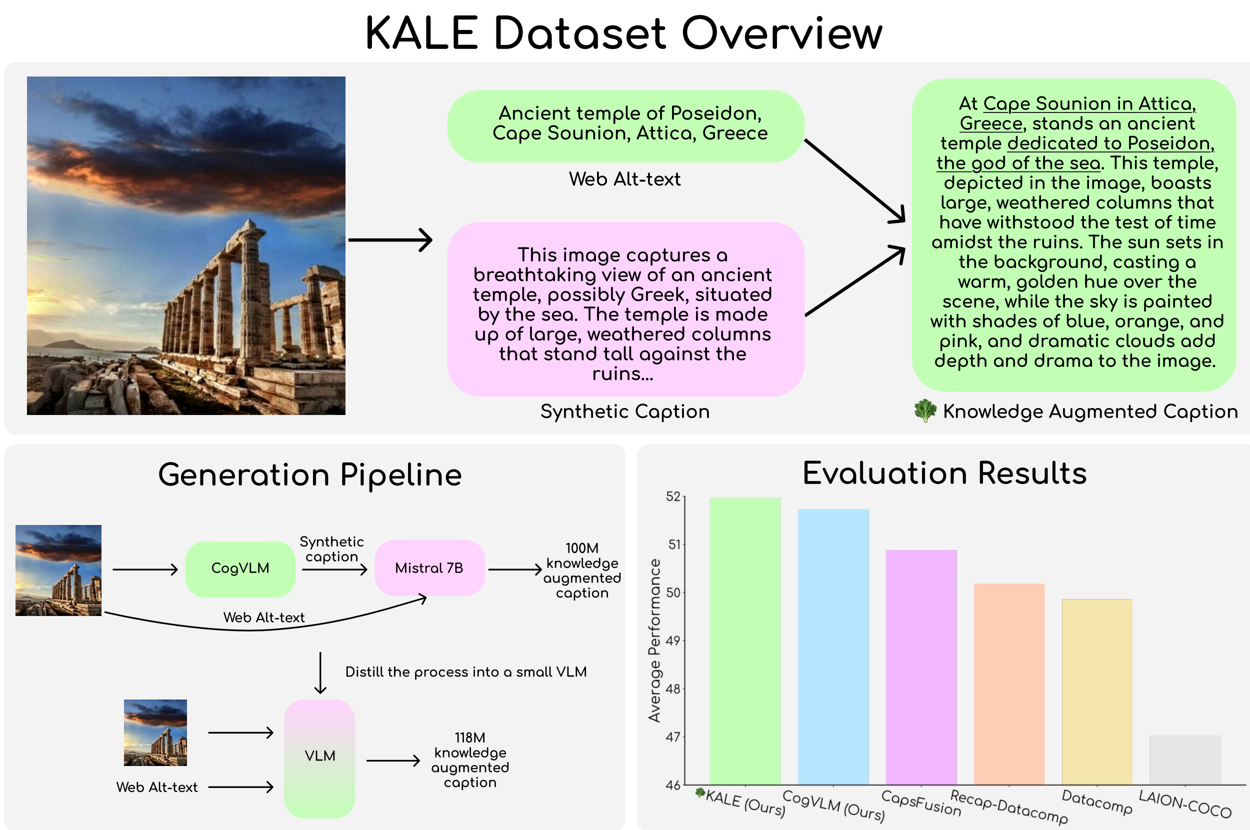

Figure 1. Overview of KALE dataset creation and performance. **Top:** Example showing how KALE combines web alt-text with synthetic captions to produce knowledge-rich dense captions. **Bottom left:** Two-stage generation pipeline for KALE, using CogVLM and Mistral to create an initial set of knowledge augmented captions, followed by training a distilled VLM to scale up to 218M samples. **Bottom right:** Evaluation results comparing KALE's average performance against popular synthetic image-text datasets.

Dataset	Scale (# of samples)	Density (avg. words/caption)	Knowledge-augmented?	Captioner size (params)
LAION-COCO[1]	600M	8.99	✗	0.5B
ReCap-Datacomp-1B [16]	1.28B	49.43	✗	7B
CapsFusion [29]	120M	22.74	✓	0.5B
KALE	218M	67.26	✓	17B (stage 1) $\rightarrow$ 2B (stage 2)

Table 1. **Comparison of open-source synthetic image-text datasets:** We compare various datasets in terms of scale (number of samples), density (average number of words per sample), whether they are knowledge-augmented (meaning that the caption includes information found in image's web scraped alt-text), and the size of the captioning model used to generate the descriptions. For KALE, we create an initial pool of 100M captions from a 17B parameter model and use it to distill a 2B parameter model that matches the performance of the larger 17B model.

2.3. Removing pipeline artifacts

Artifacts from the prompts passed into LLMs/VLMs to generate KALE occasionally leak into the generated captions. We present an example of these artifacts in Figure 3 where the system prompt to the LLM used in the rewriting stage has leaked into the generated caption. To remove these artifacts, we create a set of words that commonly appear in these artifacts such as 'real-world' or 'sentence structure' and remove sentences that contain these keywords.

3. Experiments

We validate the effectiveness of KALE by training VLMs. In this section, we outline our training and evaluation setup and present results when training on KALE.

[1]https://laion.ai/blog/laion-coco/

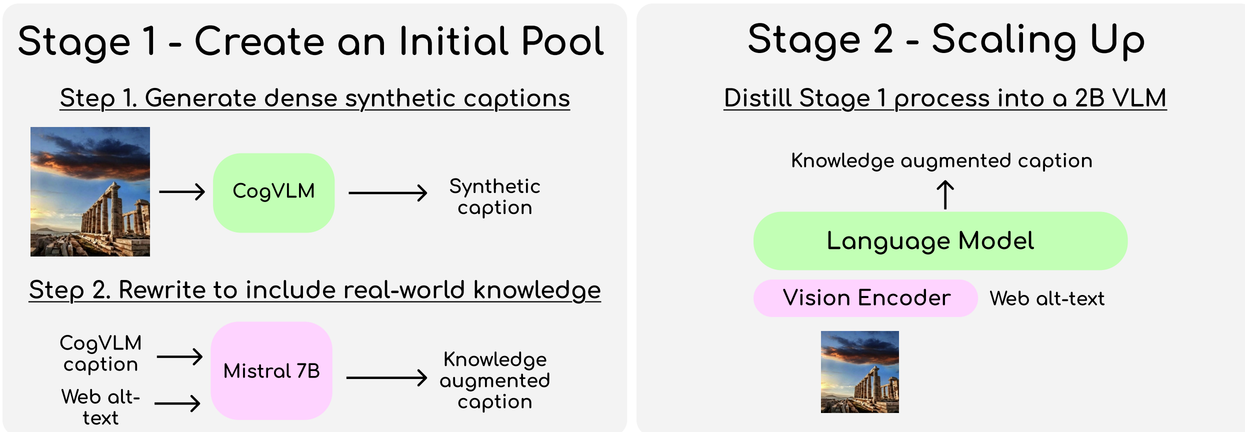

Figure 2. We generate KALE in a two stage process. **Stage 1:** We first create an initial pool of 100M knowledge-augmented dense captions using CogVLM-17B to generate dense captions for Datacomp-1B images and then augmenting these captions with real world knowledge by prompting Mistral. **Stage 2:** We use the knowledge-augmented captions from Stage 1 to train a VLM that takes image patch embeddings and Datacomp-1B captions as inputs and outputs knowledge-augmented captions. This VLM is then used to efficiently caption 118M more images from Datacomp-1B.

The Quality and project coordinator position at Credit Agricole(Dublin), **as indicated in Sentence 1,** is a part of a VIE (Venture in Ireland), and **my analysis now turns to the visual representation of this organization.** The logo, **as depicted in Sentence 2,** features a stylized green 'A' letter, which is accentuated by blue and red lines on the right side. The dynamic and modern appearance is achieved by intertwining the 'A' with the blue and red line. A horizontal green line lies beneath the letter 'A'.

Figure 3. Example of pipeline artifacts in a caption. The highlighted texts show the phrases that have leaked from the system prompt into the final output.

3.1. Training setup

We follow the Llava architecture in using a linear layer to project image patch embeddings from a vision encoder into the text embedding space. We use Qwen2.5-1.5B [28] for the language model, and SigLIP ViT-L 384 [30] for the vision encoder. We use a batch size of 80 image-text pairs and a peak learning rate of $5e^{-5}$. We train all of our models on two million samples from image-text data. We then fine-tune the model, using a peak learning rate of $3e^{-5}$, on one million multimodal instruction tuning samples from the Cauldron [12] dataset. We remove multi-image samples from the Cauldron data and sample different subsets according to the ratios used in Idefics2.

Model	Average
KALE (stage 1 only)	51.53
KALE	51.96

Table 2. Average performance shows little difference between training on stage 1 captions and a mixture of stage 1 and stage 2 (i.e. KALE).

3.2. Evaluation setup

We evaluate the instruction-tuned model on various vision-language benchmarks including TextVQA (val set) [25], VQAv2 (val lite) [1], ScienceQA [20], AI2D [9], MM-Bench [18], ChartQA [21], InfoVQA [22], OCRBench [19], RealWorldQA[1], and MMStar [4] using the lmms-eval framework [31]. This comprehensive evaluation suite covers a wide range of capabilities from general visual question answering to specialized tasks involving scientific reasoning and OCR-based comprehension.

3.3. Results

We find that pre-training on KALE captions improves downstream model performance on most VLM benchmarks, achieving the highest average performance at 51.96%. In particular, KALE shows strong performance on TextVQA (59.92%), VQAv2 (70.10%), and ScienceQA (72.68%). CogVLM's synthetic captions also demonstrate robust performance. Both KALE and CogVLM significantly outperform Datacomp-1B's noisier alt-text captions, which achieve lower scores across most benchmarks (49.86% average).

[1] https://x.ai/blog/grok-1.5v

Model	Benchmarks										
	TextVQA	VQAv2	ScienceQA	AI2D	MMBench	ChartQA	InfoVQA	OCRBench	RealWorldQA	MMStar	Avg
KALE (Ours)	**59.92**	**70.10**	**72.68**	**65.64**	58.59	23.28	29.28	**43.80**	52.42	**43.91**	**51.96**
CogVLM (Ours)	59.74	69.42	70.30	65.35	**61.60**	**23.64**	**29.53**	**43.80**	52.03	41.90	51.73
CapsFusion	57.62	67.30	71.79	62.27	60.82	22.28	27.67	43.10	52.03	**43.91**	50.88
Recap-Datacomp	58.49	67.36	71.19	63.31	52.75	23.08	28.45	42.20	**53.07**	41.90	50.18
Datacomp	57.40	67.22	69.51	61.82	59.45	22.28	28.53	42.20	50.46	39.70	49.86
LAION-COCO	54.12	65.26	65.94	59.10	55.58	21.60	26.81	38.90	44.05	38.90	47.03

Table 3. **Downstream performance:** To measure the quality of KALE in comparison to other datasets, we evaluate the instruction-tuned models across vision-language tasks. KALE maintains a slight edge in overall performance, while our CogVLM synthetic captions shows strong performance in tasks like MMBench. Both subsets of our KALE data outperform existing synthetic image-text datasets.

Earlier attempts at knowledge integration, such as CapsFusion (50.88% average), while showing improvements over the Datacomp baseline, did not achieve the same level of performance as our approach. The LAION-COCO dataset, constrained by both vocabulary size and caption density, performs the lowest at 47.03% average. Furthermore, Table 2 compares the performance of stage 1 captions generated by CogVLM and Mistral-7B with the complete KALE dataset, which combines these stage 1 captions with those from our distilled captioning model (stage 2). The combined stage 1 and 2 captions achieve performance comparable to the stage 1 captions generated by the significantly larger CogVLM model, demonstrating the effectiveness of our distilled captioning approach.

In "The Ride for Liberty- The Fugitive Slaves," painted by Eastman Johnson around 1862, a family of three, consisting of a man, woman, and child, is depicted in an oil painting on board. Measuring 21 15/16 x 26 1/8 inches, this masterpiece is housed at The Brooklyn Museum as a gift from Miss Gwendolyn O. L. Conkling. The scene unfolds with the family in mid-gallop on their horse, the man holding the reins, the woman seated behind him, and the child on his lap. The horse's tail flows behind, and the landscape, though vast and barren, is characterized by an overcast sky, hinting at a gloomy or early morning setting. The painting's style is realistic, meticulously capturing the intricacies of the subjects and the environment.

"In the intricately detailed watercolor painting by Emma Howell, dated 2021, a captivating bath landscape unfolds. The scene portrays a serene body of water, be it a lake or a river, encircled by verdant hills and lush greenery. To the left, an intriguing structure emerges, hinting at a rich history with its aged appearance. Whether it's a ruin, an old church, or a castle, the structure adds depth and intrigue to the painting. The sky above is a breathtaking canvas of blue and yellow hues, reflecting the tranquility of either dawn or dusk. The artist's signature, a testament to the work's authenticity, is discreetly placed at the bottom of the painting."

"A waterproof USB charger features an inline fuse for overload protection, as depicted in this image. The fuse, labeled as such, is a 23.7-inch long component with a built-in design. The image provides a detailed view of the fuse, showcasing its connectors and the fuse itself. The connectors are color-coded, with one being red and the other black, and the fuse is securely housed in a black casing.",

In the set of the Top Gun sequel, Tom Cruise is captured sitting on a sleek, black motorcycle with a sporty design, shiny finish, and clear windshield. He dons a green jacket adorned with an American flag patch on the left arm and wears sunglasses. The image's background reveals blurred figures and greenery, suggesting an open area or park setting.

Figure 4. More samples from KALE.

4. Related Works

KALE builds on many large-scale image-text datasets such as LAION-5B [24], Datacomp-1B [7], COYO-700M [3], and many more. These datasets were sourced from large amounts of images paired with alt-text captions found in the HTML image tags associated with these images. As LLM/VLMs have become more capable, many works have explored generating synthetic multimodal training data. There is a line of work that seeks to improve image-text datasets by using VLMs to generate synthetic captions [13, 14, 16, 23]. Works such as LaCLIP [5] took an alternative approach of rewriting the existing alt-text caption using an LLM to improve the quality of the caption. Moreover, works such as LLaVA [17] and LLaVAR [32] have synthetically generated visual question-answer pairs in the context of instruction tuning. Additionally, the BLIP-3 [27] model leverages our KALE dataset to improve the quality of their caption data.

Previous work has pointed out that synthetic captions lack real-world knowledge, limiting their applicability in many domains. CapsFusion [29] addresses this issue by augmenting LAION-COCO synthetic captions with alt-text from the LAION dataset. VeCLIP [10] also addresses this issue, but instead of using existing captions, it generates synthetic captions using a LLaVA model.

An adjacent line of work improves the text quality of multimodal data by instead sourcing web-scale interleaved image/text samples from web documents as opposed to HTML alt-text captions. Works such as MMC4 [33], OBELISC [11], MINT-1T [2], and OmniCorpus [15] all build multimodal interleaved datasets, which is a promising direction to attain high-quality and knowledge-rich multimodal data.

5. Limitations and conclusion

KALE represents a step forward in bridging the gap between descriptive synthetic captions and factual web-scale alt-text. Our experiments demonstrate that models trained on KALE consistently outperform baseline models across various benchmarks. While KALE performs favorably compared to other open-source image-text datasets, the data could still contain hallucinations, particularly in text-dense images. Future work should scale KALE to billions of image-text pairs, explore more sophisticated knowledge augmentation techniques, and investigate its impact on a broader range of multimodal tasks.

References

[1] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Vqa: Visual question answering. *International Journal of Computer Vision*, 123:4 – 31, 2015. 3

[2] Anas Awadalla, Le Xue, Oscar Lo, Manli Shu, Hannah Lee, Etash Kumar Guha, Matt Jordan, Sheng Shen, Mohamed Awadalla, Silvio Savarese, Caiming Xiong, Ran Xu, Yejin Choi, and Ludwig Schmidt. Mint-1t: Scaling open-source multimodal data by 10x: A multimodal dataset with one trillion tokens. *ArXiv*, abs/2406.11271, 2024. 4

[3] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset`, 2022. 4

[4] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. 3

[5] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. In *NeurIPS*, 2023. 4

[6] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023. 1

[7] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023. 4

[8] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. 1

[9] Aniruddha Kembhavi, Michael Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. *ArXiv*, abs/1603.07396, 2016. 3

[10] Zhengfeng Lai, Haotian Zhang, Wentao Wu, Haoping Bai, Aleksei Timofeev, Xianzhi Du, Zhe Gan, Jiulong Shan, Chen-Nee Chuah, Yinfei Yang, and Meng Cao. Veclip: Improving clip training via visual-enriched captions. *ArXiv*, 2023. 4

[11] Hugo Laurenccon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. Obelisc: An open web-scale filtered dataset of interleaved image-text documents. *ArXiv*, abs/2306.16527, 2023. 4

[12] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *ArXiv*, abs/2405.02246, 2024. 3

[13] Bo Li, Hao Zhang, Kaichen Zhang, Dong Guo, Yuanhan Zhang, Renrui Zhang, Feng Li, Ziwei Liu, and Chunyuan Li. Llava-next: What else influences visual instruction tuning beyond data?, 2024. 4

[14] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022. 4

[15] Qingyun Li, Zhe Chen, Weiyun Wang, Wenhai Wang, Shenglong Ye, Zhenjiang Jin, Guanzhou Chen, Yinan He, Zhangwei Gao, Erfei Cui, Jiashuo Yu, Hao Tian, Jiasheng Zhou, Chaochao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Pei Chu, Yi Wang, Min Dou, Changyao Tian, Xizhou Zhu, Lewei Lu, Yushi Chen, Jun-Jian He, Tong Lu, Yali Wang, Limin Wang, Dahua Lin, Yu Qiao, Bo Shi, Conghui He, and Jifeng Dai. Omnicorpus: A unified multimodal corpus of 10 billion-level images interleaved with text. *ArXiv*, abs/2406.08418, 2024. 4

[16] Xianhang Li, Haoqin Tu, Mude Hui, Zeyu Wang, Bingchen Zhao, Junfei Xiao, Sucheng Ren, Jieru Mei, Qing Liu, Huangjie Zheng, Yuyin Zhou, and Cihang Xie. What if we recaption billions of web images with llama-3? *ArXiv*, abs/2406.08478, 2024. 2, 4

[17] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *ArXiv*, abs/2304.08485, 2023. 1, 4

[18] Yuanzhan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? *ArXiv*, abs/2307.06281, 2023. 3

[19] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xucheng Yin, Cheng lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: On the hidden mystery of ocr in large multimodal models, 2024. 3

[20] 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 *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 3

[21] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq R. Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *ArXiv*, abs/2203.10244, 2022. 3

[22] Minesh Mathew, Viraj Bagal, Rubèn Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and C.V. Jawahar. Infographicvqa.

2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 2582–2591, 2021. 3

[23] Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. Improving multimodal datasets with image captioning. *ArXiv*, abs/2307.10350, 2023. 4

[24] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. *ArXiv*, abs/2210.08402, 2022. 4

[25] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019. 3

[26] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models. *ArXiv*, abs/2311.03079, 2023. 1

[27] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, Shrikant B. Kendre, Jieyu Zhang, Can Qin, Shu Zhen Zhang, Chia-Chih Chen, Ning Yu, Juntao Tan, Tulika Awalgaonkar, Shelby Heinecke, Huan Wang, Yejin Choi, Ludwig Schmidt, Zeyuan Chen, Silvio Savarese, Juan Carlos Niebles, Caiming Xiong, and Ran Xu. xgen-mm (blip-3): A family of open large multimodal models. *ArXiv*, 2024. 4

[28] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Ke-Yang Chen, Kexin Yang, Mei Li, Min Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yunyang Wan, Yunfei Chu, Zeyu Cui, Zhenru Zhang, and Zhi-Wei Fan. Qwen2 technical report. *ArXiv*, abs/2407.10671, 2024. 1, 3

[29] Qiying Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Xinlong Wang, and Jingjing Liu. Capsfusion: Rethinking image-text data at scale. *ArXiv*, abs/2310.20550, 2023. 1, 2, 4

[30] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11941–11952, 2023. 3

[31] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models, 2024. 3

[32] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tongfei Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *ArXiv*, abs/2306.17107, 2023. 4

[33] Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *ArXiv*, abs/2304.06939, 2023. 4