ARGUS: Vision-Centric Reasoning with Grounded Chain-of-Thought

Supplementary Material

	Re-encoding		Re-sampling	
Method	ChartQA	V-Star	ChartQA	V-Star
0 (*Original*)	73.1	67.3	**73.9**	**67.0**
+ 20%	**73.3**	67.5	73.3	66.5
+ 40%	72.7	**68.1**	73.1	64.4
+ 60%	70.1	66.5	72.5	63.3
+ 80%	70.4	64.5	71.2	60.7

Table A. Impact of RoI context expansion ratios on the performance of re-encoding and re-sampling strategies, evaluated on the ChartQA [55] and V-Star [96] benchmarks. Re-encoding demonstrates improved performance with larger context regions, while re-sampling favors the original bounding box size.

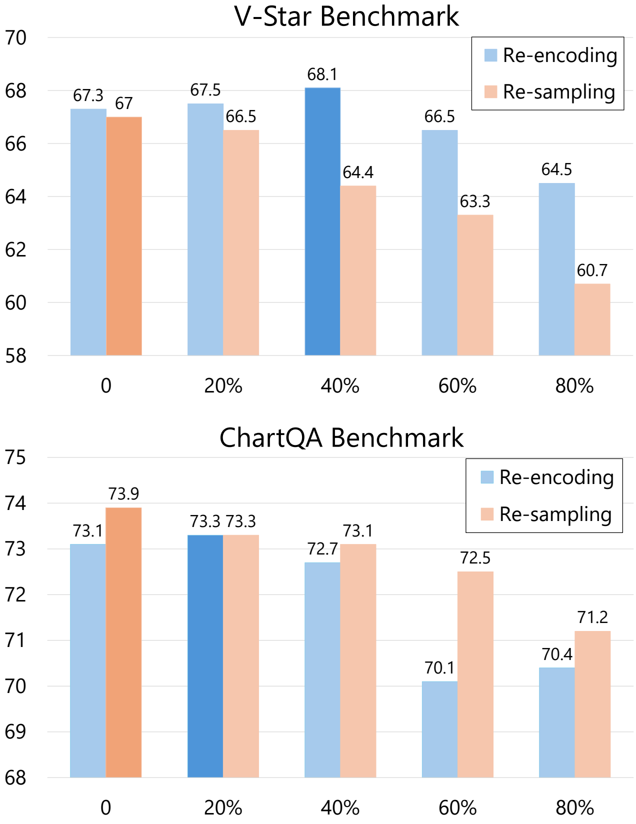


Figure A. Performance comparison of re-encoding and re-sampling strategies under varying region context expansion ratios. Re-encoding achieves optimal performance with an expanded context region (20% to 40% expansion), while re-sampling performs best with the original box size (0% expansion). The optimal performance points for each strategy are highlighted in darker colors.

A. Additional Experiments and Analysis

A.1. RoI Context Expansion

To investigate the impact of region of interest (RoI) context expansion on Argus, we examined how expanding a predicted bounding box affects performance. Specifically, we expanded the bounding box by a fixed ratio to include additional context around the predicted center. If the expanded region exceeded the image boundaries, it was cropped to fit within them. Table A and Figure A present the performance evaluation of various expansion ratios using two distinct visual re-engagement strategies. Our results reveal the following insights.

Re-encoding Strategy. The re-encoding approach benefits from a moderate expansion of the context region. Optimal performance is achieved with a 20 to 40% expansion ratio on the ChartQA [55] and V-Star [96] benchmarks. The additional context helps mitigate issues stemming from overly tight or slightly inaccurate bounding boxes, which are common in object grounding tasks. Moreover, the larger context aids in localizing the bounding box's relative position within the image, which is particularly beneficial for tasks that require both local and global context reasoning.

Re-sampling Strategy. Unlike re-encoding, the re-sampling strategy method achieves its best performance with the original bounding box size. This can be attributed to an inherent context-expansion mechanism that leverages overlapping patches, utilizing all patch embeddings that intersect with the bounding box region as the input to the re-engagement module. As a result, further expansion of the bounding box does not yield additional benefits.

Effect of Excessive Expansion. For both strategies, overly large context regions hurt performance. Including excessive irrelevant information introduces noise, which distracts the model from focusing on the most relevant RoIs. This counteracts the advantages of RoI grounding and diminishes overall effectiveness.

Choice of LLMs and Architectural Designs. In Table B, we include Vicuna 8B, 13B, and Llama 8B as LLM backbones on Argus for comparison. The results show that our method generalizes to different LLMs. In addition, stronger LLMs (in size and data) lead to direct gains in performance.

Multi-RoI Scenarios. The CoT reasoning instruction tuning datasets that we used in our work generally follow a single-RoI setting. In Table C, we extend Argus to multi-RoI reasoning with a simple multi-step framework. First, we (1) prompt the model to output RoIs (objects) about the questions in the text format. Then (2) after parsing,

Model	Vision-Centric Tasks						Text Understanding					General Tasks				
Method	Avg	V-Star	CV-Bench^{2D}	CV-Bench^{3D}	MMVP	RealworldQA	Avg	ChartQA	OCRBench	TextVQA	DocVQA	Avg	MMMU^{V}	MMB	SEED^{I}	GQA
Experiments of Argus on other LLMs and Model Sizes																
Vicuna-7B [14]	57.5	64.9	61.4	59.3	41.7	60.1	65.3	70.5	52.3	67.5	71.0	61.8	38.8	69.2	75.3	63.9
Vicuna-13B [14]	60.2	66.5	64.4	62.7	43.4	64.2	70.2	74.6	56.9	74.2	75.1	63.3	39.9	72.5	75.9	65.1
Llama-8B [81]	62.2	68.1	68.5	64.2	45.5	64.6	70.1	74.8	56.7	73.6	75.4	63.6	40.4	72.9	75.8	65.1

Table B. Argus supports various choices of LLM backbones. A larger and stronger backbone generally leads to better visual reasoning performance.

Model	V-Star	CV-Bench^{3D}
Argus (single-RoI)	68.1	64.2
Argus (multi-RoI)	**78.5**	**69.6**

Table C. Extension to multiple RoI show great improvement in vision-centric benchmarks [85, 96] where .

we conduct CoT reasoning for each separate object, and finally (3) merge multiple CoT grounding boxes and textual thoughts into one joint CoT signal for question answering. The results in the table show that this extension *improves the performance by large margins* on two visual reasoning benchmarks [85, 96], demonstrating the flexibility of Argus in handling multi-RoI settings.

Discussion about Performance Discrepancy. The performance discrepancy in CV-Bench^{3D} is likely due to the bias of the multi-RoI data, which is not extensively covered in our training data. As a result, the results of the multi-RoI extension experiment shown in Table C demonstrate significant performance increase. For the performance discrepancy of MMMU [105] and GQA [25], it is caused by strong language biases, as explained in [85] (Sec. 3.1). These two benchmarks depend more on language cues rather than visual input to correctly answer the questions, and thus we believe that a more language-oriented training data curation can lead to better performance.

B. Additional Experiment Details

B.1. Visual Foundation Models and LLMs

CLIP [65]. CLIP learns a unified embedding space for visual and textual content through contrastive learning. It optimizes the alignment between matching image-caption pairs while simultaneously pushing apart non-matching pairs in the embedding space. Due to its robust cross-modal understanding capabilities, it has established itself as the predominant vision encoder for multimodal large language models (MLLMs). In our implementation, we leverage the official huggingface checkpoint[1] of the ViT-L/14 architecture to initialize our CLIP vision expert. Following [72], we interpolate the positional embedding to obtain the input image dimension 448 × 448.

ConvNeXt [52]. ConvNeXt represents a modern evolution of convolutional neural networks (CNNs) that bridges the gap between CNNs and transformers. By incorporating transformer-inspired design principles while preserving the inherent advantages of convolutional architectures, it achieves exceptional performance across diverse vision tasks, making it an excellent choice as a vision expert. We employ the official checkpoint[2] of a ConvNeXt-XXLarge model, which has been pre-trained on LAION-2B and fine-tuned on ImageNet-1K. The input image dimension is set to 1024 × 1024.

EVA-02 [18, 19]. EVA-02 is a vision foundation model that achieves superior performance with moderate model sizes. It incorporates Transformer architecture designs and utilizes masked image modeling pre-training with features from a large CLIP vision encoder. For this work, we specifically employ the EVA-02-L/16 model checkpoint pre-trained on detection-focused datasets including COCO [44] and Objects365 [71], making it particularly well-suited for perceptive tasks. We utilize the official checkpoint[3] and process input images at a resolution of 1024 × 1024.

Llama 3 [81]. Llama 3 represents the latest advancement in open-sourced large language models (LLMs), incorporating significant improvements over its predecessors in instruction-following capabilities and reasoning ability. For our implementation, we employ the official checkpoint[4] of the Meta-Llama-3-8B-Instruct model, which has been specifically fine-tuned for instruction-following scenarios.

B.2. Implementation Details

Global Training Hyperparameters. Table D details the global training hyperparameters we have employed across

[1] https://huggingface.co/openai/clip-vit-large-patch14-336

[2] https://huggingface.co/timm/convnext_xxlarge.clip_laion2b_soup_ft_in1k

[3] https://huggingface.co/Yuxin-CV/EVA-02/blob/main/eva02/det/eva02_L_coco_det_sys_o365.pth

[4] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

Parameters	Stage 1	Stage 2
Learning rate	$1e^{-3}$	$2e^{-5}$
Vision encoders	trainable	trainable
Projector	trainable	trainable
LLM backbone	frozen	trainable
Global batch size	256	256
Optimizer	AdamW	AdamW
Weight decay	0.0	0.0
Beta coefficient β_1	0.9	0.9
Beta coefficient β_2	0.999	0.999
Epsilon coefficient ϵ	$1e^{-8}$	$1e^{-8}$
Gradient accumulation steps	1	1
Warmup ratio	0.03	0.03
Epochs	1	1
Projector type	mlp2×	mlp2×
Learning rate scheduler	cosine	cosine
Gradient checkpointing	true	true
Precision	bfloat16	bfloat16
Max sequence length	2048	3072

Table D. Global training hyperparameters of stage 1 pre-training and stage 2 supervised fine-tuning (SFT) for Argus.

both stages of Argus training. For stage 1, we initialize the vision experts using pre-aligned checkpoints as described in Eagle [72], while the MLP projector is randomly initialized. In stage 2, both the vision experts and MLP projectors are initialized using the checkpoints obtained from stage 1. The pre-training stage is trained with 32× NVIDIA A100 GPUs for 4 hours, , while the supervised fine-tuning (SFT) stage utilizes 64× NVIDIA A100 GPUs and requires 28 hours of training.

Vision Encoder Hyperparameters. Table E presents the specific hyperparameters for each vision encoder integrated into our model. Each encoder is optimized for different input image resolutions and operates with distinct hidden feature dimensions. Following feature extraction, we resize the spatial dimensions of all feature embeddings to 32×32 and concatenate them along the feature channel dimension, producing a unified tensor of shape $32 \times 32 \times 5120$ tensor. This concatenated representation is then processed by the multimodal MLP projector, which maps the feature channels to match the LLM's hidden dimension of 4096, ultimately generating 1024 visual tokens.

B.3. Training Dataset

In this section, we detail the datasets utilized in our model training pipeline. For pre-training, we adopt the standard LLaVA-595K [48] dataset, following the training protocols established by recent state-of-the-art MLLMs. For supervised fine-tuning, we employ a diverse mixture of datasets from multiple sources.

Parameters	Values
CLIP input resolution	448×448
CLIP hidden size	1024
ConvNeXt input resolution	1024×1024
ConvNeXt hidden size	3072
EVA-02 input resolution	1024×1024
EVA-02 hidden size	1024
Grid size	32×32
Aspect ratio	square
Pre-processing	padding & resizing
Global hidden size	5120

Table E. Hyperparameters for vision encoder designs of Argus.

Eagle-1.8M [72]. Eagle-1.8M represents a comprehensive collection of conversational data aggregated from various specialized datasets, comprising LLaVA-Instruct [48] (665K), DocVQA [56] (39K), synDog-EN [33] (50K), ChartQA [55] (28K), DVQA [27] (25K), AI2D [32] (15K), ShareGPT-4V [9] (100K), LAION-GPT4v [80] (11K), LVIS-Instruct4V [89] (220K), LRV-Instruct [47] (150K), Geo170K [21] (120K), LLaVAR [109] (20K), Visual7W [115] (70K), and Open-Hermes 2.5 [84] (300K). This diverse collection spans a wide spectrum of reasoning scenarios and establishes a robust foundation for our vision-centric reasoning capabilities.

VCoT [70]. VCoT comprises a diverse collection of datasets featuring paired bounding box annotations and image-question pairs, including TextVQA[74] (16K), TextCaps [73] (32K), DocVQA [56] (33K), DUDE [87] (15K), SROIE [24] (4K), Birds-200-2011 [88] (10K), Flickr30K [64] (136K), Visual7W [115] (43K), InfographicsVQA [57] (15K), VSR [46] (3K), GQA [25] (88K), and Open images [84] (43K).

GRIT [63]. Grounded Image-Text pairs (GRIT) is a large-scale dataset extracted from COYO-700M [4] and LAION-2B [80]. It is constructed through a pipeline that extracts and links noun phrases and referring expressions in image captions to their corresponding visual regions. Each sample contains an image, caption, extracted noun chunks with corresponding bounding boxes, and two CLIP scores [65] (from ViT-B/32 and ViT-L/14) measuring text-image similarity. Following [42], we retain 756K samples after filtering out entries with CLIP scores below 0.35.

Shikra [8]. Shikra offers a curated collection of perception-centric datasets specifically designed for object grounding instruction tuning. From its composition, we utilize 326K visual grounding-oriented samples from the RefCOCO-family datasets (RefCOCO, RefCOCO+, RefCOCOg)[31, 103], Visual Genome [37], Visual-7w [115], and Flickr30K [64].

C. Additional Qualitative Visualization

We provide additional visualization of the qualitative results on multimodal benchmarks [55, 96] in Figure B and Figure C.

D. Limitations and Future Work

While we have made substantial progress in exploring the design space of MLLMs for vision-centric reasoning tasks, we acknowledge several limitations in our current approach. This section discusses these limitations and outlines potential directions for future research.

Model Capacity. Our investigation primarily focuses on the design space of visual CoT mechanisms with grounding signals, utilizing the 8-billion parameter Llama3 [81] model as our LLM decoder backbone. This specific architectural choice may limit the generalizability of our findings. A natural extension of this work would be to evaluate our approach across a spectrum of model scales to validate whether our findings remain consistent in larger architectural configurations.

Dataset Complexity. While Argus leverages a diverse combination of multimodal reasoning, visual CoT, and perception/grounding datasets, the current landscape of visual CoT signals remains limited in diversity. Unlike language-based CoT signals, which are abundantly available in internet-scale text corpora and existing language datasets, visual CoT signals are rarely present in large-scale vision-language datasets. Although we have demonstrated significant improvements with the available data, we acknowledge that access to larger-scale, higher-quality visual CoT data would likely yield substantial performance gains and potentially reveal novel emergent capabilities. We believe such data could be derived from existing visual perception datasets or through targeted human annotation efforts, presenting an important avenue for future research.

Expanded Vision-centric Tasks Coverage. While our evaluation of Argus focuses on visual question answering and referring object grounding tasks – which effectively demonstrate the synergy between perception and reasoning objectives – we recognize this scope as potentially not comprehensive. In pursuit of developing a truly vision-centric generalist model, a crucial capability would be support for open-world detection tasks. However, given the substantial computational resources and time required for such an investigation, we regard this exploration as future work.

References

[1] 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*, 2023. 2, 5, 6, 7

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6

[3] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of thoughts: Solving elaborate problems with large language models. In *AAAI*, 2024. 3

[4] 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. 3

[5] Nadine Chang, Zhiding Yu, Yu-Xiong Wang, Animashree Anandkumar, Sanja Fidler, and Jose M Alvarez. Image-level or object-level? A tale of two resampling strategies for long-tailed detection. In *ICML*, 2021. 4

[6] Hong-You Chen, Zhengfeng Lai, Haotian Zhang, Xinze Wang, Marcin Eichner, Keen You, Meng Cao, Bowen Zhang, Yinfei Yang, and Zhe Gan. Contrastive localized language-image pre-training. *arXiv preprint arXiv:2410.02746*, 2024. 2

[7] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. MiniGPT-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 6, 7

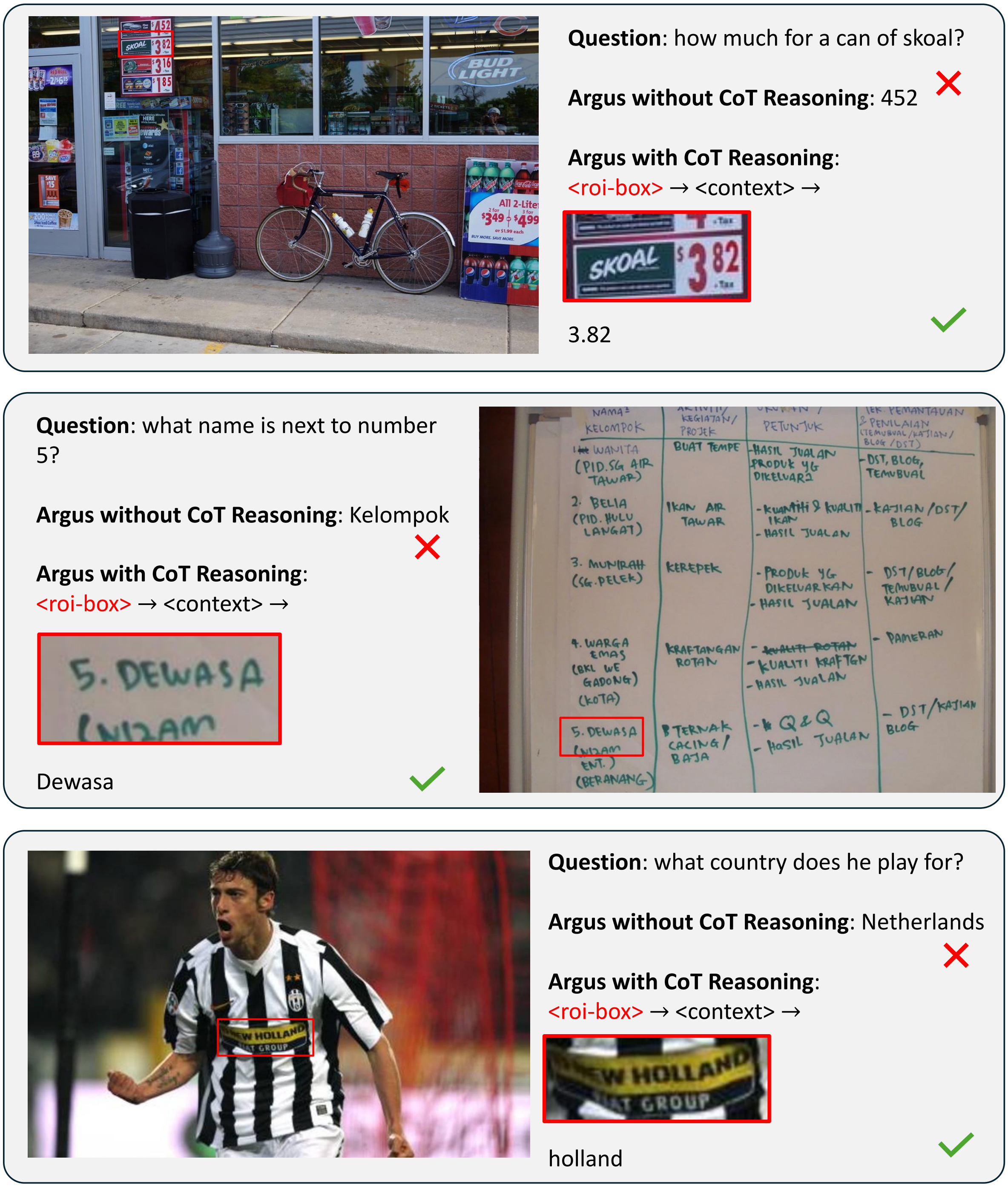
[8] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal LLM's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 4, 5, 6, 7, 3

[9] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. ShareGPT4v: Improving large multi-modal models with better captions. In *ECCV*, 2024. 5, 3

[10] Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *TMLR*, 2023. 3

[11] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: Universal image-text representation learning. In *ECCV*, 2020. 6, 7

[12] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. How far are we to GPT-4V? closing the gap to commercial multimodal models



Question: how much for a can of skoal?

Argus without CoT Reasoning: 452 ✗

Argus with CoT Reasoning:
<roi-box> → <context> →

3.82 ✓

Question: what name is next to number 5?

Argus without CoT Reasoning: Kelompok ✗

Argus with CoT Reasoning:
<roi-box> → <context> →

Dewasa ✓

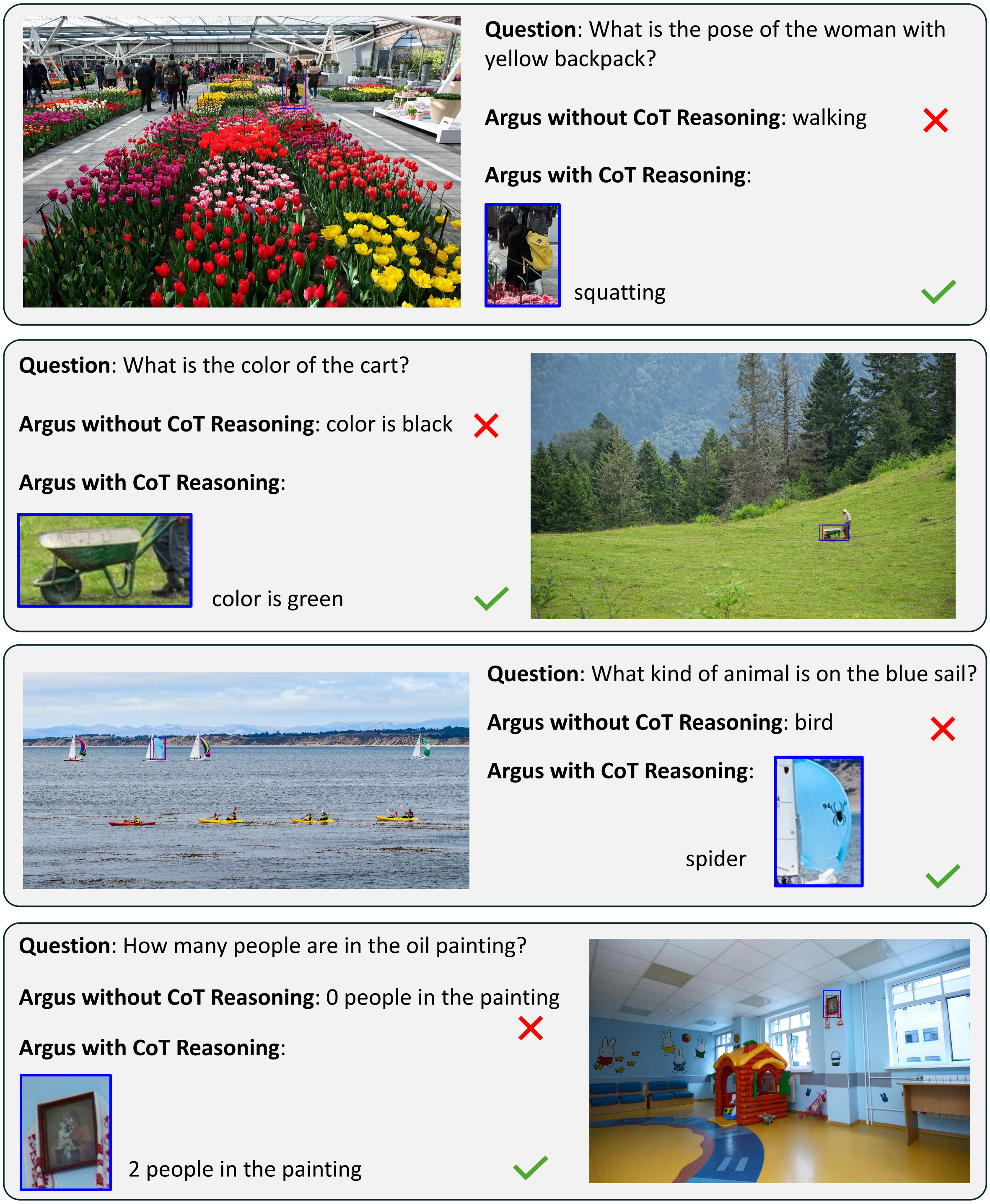
Question: what country does he play for?

Argus without CoT Reasoning: Netherlands ✗

Argus with CoT Reasoning:
<roi-box> → <context> →

holland ✓

Figure B. Argus performance on TextVQA [74] benchmark, emphasizing on text localization and interpretation in the images.



Question: What is the pose of the woman with yellow backpack?

Argus without CoT Reasoning: walking ✗

Argus with CoT Reasoning:

squatting ✓

Question: What is the color of the cart?

Argus without CoT Reasoning: color is black ✗

Argus with CoT Reasoning:

color is green ✓

Question: What kind of animal is on the blue sail?

Argus without CoT Reasoning: bird ✗

Argus with CoT Reasoning:

spider ✓

Question: How many people are in the oil painting?

Argus without CoT Reasoning: 0 people in the painting ✗

Argus with CoT Reasoning:

2 people in the painting ✓

Figure C. Argus performance on V-Star [96] benchmark, emphasizing visual perception of objects and regions in complex scenarios. Ground truth bounding boxes are represents in red, and our predicted bounding boxes are in blue.

with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 2

[13] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024. 2, 6

[14] 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. *https://vicuna. lmsys. org*, 2023. 3, 2

[15] Maurizio Corbetta and Gordon L Shulman. Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience*, 2002. 1

[16] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. TransVG: End-to-end visual grounding with transformers. In *ICCV*, 2021. 6, 7

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3

[18] 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 *CVPR*, 2023. 3, 2

[19] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA-02: A visual representation for neon genesis. *Image and Vision Computing*, 2024. 3, 5, 2

[20] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In *NeurIPS*, 2020. 6, 7

[21] Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and Lingpeng Kong. G-LLaVA: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*, 2023. 5, 3

[22] Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, Lewei Lu, Tong Lu, Yu Qiao, Jifeng Dai, and Wenhai Wang. Mini-InternVL: a flexible-transfer pocket multi-modal model with 5% parameters and 90% performance. *Visual Intelligence*, 2(1):1–17, 2024. 2

[23] Jiaxin Ge, Hongyin Luo, Siyuan Qian, Yulu Gan, Jie Fu, and Shanghang Zhang. Chain of thought prompt tuning in vision language models. *arXiv preprint arXiv:2304.07919*, 2023. 3

[24] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. ICDAR2019 competition on scanned receipt ocr and information extraction. In *ICDAR*, 2019. 5, 3

[25] Drew A Hudson and Christopher D Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 2, 5, 6, 3

[26] William James. *Psychology, briefer course*. Harvard University Press, 1984. 1

[27] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. DVQA: Understanding data visualizations via question answering. In *CVPR*, 2018. 5, 3

[28] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR-modulated detection for end-to-end multi-modal understanding. In *ICCV*, 2021. 6, 7

[29] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019. 4

[30] Stephen Kaplan and Marc G Berman. Directed attention as a common resource for executive functioning and self-regulation. *Perspectives on psychological science*, 2010. 1

[31] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 2, 3

[32] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016. 5, 3

[33] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *ECCV*, 2022. 5, 6, 3

[34] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 2

[35] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *ICCV*, 2023. 2

[36] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *NeurIPS*, 2022. 3

[37] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123:32–73, 2017. 5, 3

[38] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 128(7):1956–1981, 2020. 5

[39] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and

Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 2, 3

[40] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 2

[41] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-Gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024. 1, 3, 5, 6

[42] Zejun Li, Ruipu Luo, Jiwen Zhang, Minghui Qiu, and Zhongyu Wei. VoCoT: Unleashing visually grounded multi-step reasoning in large multi-modal models. *arXiv preprint arXiv:2405.16919*, 2024. 3, 4

[43] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. VILA: On pre-training for visual language models. In *CVPR*, 2024. 1, 2, 3, 5

[44] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 2

[45] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, Jiaming Han, Siyuan Huang, Yichi Zhang, Xuming He, Hongsheng Li, and Yu Qiao. SPHINX: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023. 3, 4

[46] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. 5, 3

[47] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023. 5, 3

[48] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 2, 3, 4, 5

[49] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge, 2024. 1, 2, 3, 4, 5, 6

[50] Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, Lei Zhang, Jianfeng Gao, and Chunyuan Li. LLaVA-Plus: Learning to use tools for creating multimodal agents. *arXiv preprint arXiv:2311.05437*, 2023. 3

[51] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, 2024. 2, 4, 6, 7

[52] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022. 3, 5, 2

[53] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[54] Pan Lu, Swaroop Mishra, Tanglin 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. 3

[55] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *ACL*, 2022. 2, 5, 6, 1, 3, 4

[56] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, 2021. 2, 5, 6, 3

[57] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *WACV*, 2022. 5, 3

[58] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang, Ankur Jain, Hongyu Hè, Max Schwarzer, Tom Gunter, Xiang Kong, Aonan Zhang, Jianyu Wang, Chong Wang, Nan Du, Tao Lei, Sam Wiseman, Guoli Yin, Mark Lee, Zirui Wang, Ruoming Pang, Peter Grasch, Alexander Toshev, and Yinfei Yang. MM1: Methods, analysis & insights from multimodal LLM pre-training. *arXiv preprint arXiv:2403.09611*, 2024. 6

[59] Robert J Morecraft, Changiz Geula, and M-Marsel Mesulam. Architecture of connectivity within a cingulo-fronto-parietal neurocognitive network for directed attention. *Archives of neurology*, 1993. 1

[60] Vernon B Mountcastle. Brain mechanisms for directed attention. *Journal of the Royal Society of Medicine*, 1978. 1

[61] Minheng Ni, Yutao Fan, Lei Zhang, and Wangmeng Zuo. Visual-O1: Understanding ambiguous instructions via multi-modal multi-turn chain-of-thoughts reasoning. *arXiv preprint arXiv:2410.03321*, 2024. 3

[62] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024. 2

[63] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 5, 3

[64] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015. 5, 3

[65] 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 *ICML*, 2021. 2, 3, 5

[66] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. GLaMM: Pixel grounding large multimodal model. In *CVPR*, 2024. 3, 6

[67] Tianhe Ren, Qing Jiang, Shilong Liu, Zhaoyang Zeng, Wenlong Liu, Han Gao, Hongjie Huang, Zhengyu Ma, Xiaoke Jiang, Yihao Chen, Yuda Xiong, Hao Zhang, Feng Li, Peijun Tang, Kent Yu, and Lei Zhang. Grounding DINO 1.5: Advance the edge of open-set object detection. *arXiv preprint arXiv:2405.10300*, 2024. 4

[68] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 2

[69] Daniel Rose, Vaishnavi Himakunthala, Andy Ouyang, Ryan He, Alex Mei, Yujie Lu, Michael Saxon, Chinmay Sonar, Diba Mirza, and William Yang Wang. Visual chain of thought: bridging logical gaps with multimodal infillings. *arXiv preprint arXiv:2305.02317*, 2023. 3

[70] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual CoT: advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *NeurIPS*, 2024. 3, 4, 5, 6

[71] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, 2019. 2

[72] Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, Bryan Catanzaro, Andrew Tao, Jan Kautz, Zhiding Yu, and Guilin Liu. Eagle: Exploring the design space for multimodal llms with mixture of encoders. In *ICLR*, 2025. 1, 2, 3, 4, 5, 6

[73] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *ECCV*, 2020. 5, 3

[74] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *CVPR*, 2019. 2, 5, 6, 3

[75] Anthropic Team. Introducing the next generation of claude, 2024. 1, 2

[76] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 3

[77] Emu3 Team. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 3

[78] Gemini Team. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1, 2

[79] Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 2

[80] LAION Team. Laion-gpt4v dataset, 2023. 5, 3

[81] Meta Team. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 3, 5, 2, 4

[82] OpenAI Team. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 2, 3, 6

[83] OpenGVLab Team. InternVL2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy, 2024. 6, 7

[84] Teknium. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants., 2023. 5, 3

[85] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 1, 2, 3, 4, 5, 6

[86] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *CVPR*, 2024. 1, 2, 6

[87] Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickael Coustaty, Bertrand Anckaert, Ernest Valveny, Matthew Blaschko, Sien Moens, and Tomasz Stanislawek. Document understanding dataset and evaluation (dude). In *ICCV*, 2023. 5, 3

[88] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 5, 3

[89] Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. To see is to believe: Prompting GPT-4v for better visual instruction tuning. *arXiv preprint arXiv:2311.07574*, 2023. 5, 3

[90] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, 2022. 6, 7

[91] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2

[92] 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 preprint arXiv:2311.03079*, 2023. 3

[93] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and

Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *ICLR*, 2023. 3

[94] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022. 3

[95] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual ChatGPT: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023. 3

[96] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal LLMs. In *CVPR*, 2024. 2, 3, 6, 7, 1, 4

[97] xAI Team. Grok, 2024. 1, 2, 6

[98] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *CVPR*, 2024. 3

[99] Jiawei Yang, Katie Z Luo, Jiefeng Li, Congyue Deng, Leonidas Guibas, Dilip Krishnan, Kilian Q Weinberger, Yonglong Tian, and Yue Wang. Denoising vision transformers. In *ECCV*, 2024. 4

[100] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. UniTAB: Unifying text and box outputs for grounded vision-language modeling. In *ECCV*, 2022. 6, 7

[101] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *NeurIPS*, 2024. 3

[102] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *ICLR*, 2023. 2, 3, 4, 5, 6, 7

[103] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016. 2, 5, 6, 3

[104] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*, 2018. 6, 7

[105] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024. 2, 6

[106] Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Lei Zhang, Chunyuan Li, and Jianwei Yang. Llava-grounding: Grounded visual chat with large multimodal models. In *ECCV*, 2024. 4

[107] Haotian Zhang, Haoxuan You, Philipp Dufter, Bowen Zhang, Chen Chen, Hong-You Chen, Tsu-Jui Fu, William Yang Wang, Shih-Fu Chang, Zhe Gan, and Yinfei Yang. Ferret-v2: An improved baseline for referring and grounding with large language models. In *COLM*, 2024. 2, 4

[108] Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198*, 2024. 3

[109] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. LLaVAR: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023. 5, 3

[110] Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, and Joyce Chai. GroundHog: Grounding large language models to holistic segmentation. In *CVPR*, 2024. 3, 4

[111] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. In *ICLR*, 2023. 3

[112] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023. 3

[113] Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. PSALM: Pixelwise segmentation with large multi-modal model. In *ECCV*, 2024. 3, 4

[114] Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. DdCoT: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. In *NeurIPS*, 2023. 3

[115] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *CVPR*, 2016. 5, 3