

A Benchmark Details

InfoSeek InfoSeek is a visual question answering (VQA) dataset tailored for information-seeking questions that cannot be answered with only common sense knowledge. It combines human-annotated and automatically collected data from visual entity recognition datasets and Wikidata, providing over one million examples for model fine-tuning and validation [25]. For InfoSeek, the ground truth answers for test sets are not publicly available, so we follow prior work [28, 34, 35] and report results on the validation sets. These sets include questions not seen during training and those associated with unseen entities.

OVEN OVEN (Open-domain Visual Entity Recognition) challenges models to select among six million possible Wikipedia entities, making it a general visual recognition benchmark with the largest number of labels. It is constructed by re-purposing 14 existing datasets with all labels grounded onto one single label space: Wikipedia entities [39]. Similar with Infoseek, the ground truth answers for the test sets of OVEN are not publicly available, so we also report results on the validation sets.

MRAG-Bench MRAG-Bench is a multimodal retrieval-augmented generation benchmark designed to evaluate the performance of large vision-language models (LVLMs) in scenarios where visual knowledge retrieval is more beneficial than textual information. It consists of 16,130 images and 1,353 human-annotated multiple-choice questions across nine distinct scenarios [40].

OK-VQA OK-VQA includes more than 14,000 open-ended questions that require external knowledge to answer. The dataset is manually filtered to ensure all questions necessitate information beyond the image content, such as from Wikipedia [26].

A-OKVQA A-OKVQA is a crowdsourced visual question answering dataset composed of approximately 25,000 questions requiring a broad base of commonsense and world knowledge to answer. Unlike existing knowledge-based VQA datasets, the questions generally cannot be answered by simply querying a knowledge base and instead require some form of commonsense reasoning about the scene depicted in the image [27].

ViQuAE ViQuAE is a dataset focusing on knowledge-based visual question answering about named entities. It covers a wide range of entity types, such as persons, landmarks, and products, and evaluates models’ abilities to ground visual content with knowledge base information [41].

CVQA CVQA (Culturally-diverse Multilingual Visual Question Answering) dataset is a benchmark that offers a broad, inclusive representation by incorporating culturally-driven images and questions from a wide range of countries and languages[42]. In this study, we evaluate five of the most widely used languages in CVQA: Chinese, Russian, Spanish, Portuguese, and Bulgarian.

For all benchmarks, we follow the official evaluation protocols to compute the accuracy of the model’s responses. Specifically: (1) For InfoSeek, OK-VQA, A-OKVQA, and ViQuAE, we use exact match evaluation to verify whether the model’s response exactly matches the ground-truth answers. (2) For OVEN, we adopt the official evaluation script, which uses BM25 [57] to match the model’s answer with relevant Wikipedia entities. (3) For MRAG-Bench and CVQA, which are in multiple-choice format, we evaluate accuracy by checking whether the model selects the correct option.

B Implementation Details

• **Knowledge Retrieval** Our knowledge base is constructed using the Wikipedia-based Image-Text (WIT) dataset[31], which consists of 37.5 million curated image-text pairs from Wikipedia articles across 108 languages. Based on WIT knowledge base, we implement a CLIP-based image-to-image retrieval system to identify the most relevant external knowledge. Following the stage-1 retrieval methodology of RoRA[32], we first encode all images in WIT using a frozen CLIP image encoder[30] to build a dense vector-search database. Given a query image \mathcal{I} , its CLIP embedding $CLIP(\mathcal{I})$ is compared against all vectors in the knowledge base via cosine similarity, followed by softmax normalization over the similarity scores. The image retriever then returns the top- k highest-scoring images along with their associated textual descriptions.

587 • **Memory Encoding** Given the retrieved image-text pairs, we employ a memory encoder, consisting
588 of a VLM and a Q-Former to compress multimodal information. For Each image-text pair is
589 compressed into an 8-token vector. These token vectors are then concatenated and passed into
590 the inference-time model. For Qwen2.5-Instruct VL, we uses Qwen2.5-Instruct VL as both the
591 inference-time model and the memory encoder, and for Qwen2-Instruct VL, we uses Qwen2-Instruct
592 VL as both the inference-time model and the memory encoder.

593 • **Answer Generation** The concatenated compressed tokens are plug into the inference-time model
594 to generate answers. We should note that *our compression module is model-agnostic, allowing the*
595 *memory encoder to be plugged into other LMs*. This flexibility is further demonstrated in Section 4.3.

596 C Limitations

597 • **Evaluation Benchmarks** While we evaluate our method on 6 multimodal and 2 multilingual
598 reasoning tasks, most of benchmarks are static and synthetic. Real-world applications with dynamic
599 or noisy inputs (*e.g.*, web data, live video) may introduce challenges.

600 • **Multi-Agent Settings** Our current framework is designed and evaluated in a single-model setting,
601 where one inference language model uses the continuous memory module for enhanced reasoning.
602 However, many real-world applications involve multiple collaborating agents or a combination of
603 LMs and VLMs. Whether our continuous memory can effectively transmit and share knowledge
604 across multiple models remains unexplored and will be investigated in future work.

605 D Training Efficiency

606 To evaluate the training efficiency of
607 our method, we assess the performance
608 of CoMEM on Qwen2.5-VL using
609 the Infoseek benchmark under varying
610 amounts of training data and trainable
611 parameters. In the original setting, we
612 use only 15.6k training samples and fine-
613 tune 1.2% of the total parameters. For
614 the data variation setting, we scale the
615 training data by factors of 0.25x, 0.5x,
616 2x, and 4x. For the parameter variation
617 setting, we adjust the LoRA rank and
618 the number of Q-Former layers by the
619 same scaling factors to control the num-
620 ber of trainable parameters.

621 As shown in Table 6, increasing the
622 training data by 2x or even 4x results in
623 only marginal performance gains, suggesting that the original data size is already adequate for effective training. Similarly, increasing the number of trainable parameters does not yield improvements, while reducing them below the original configuration leads to a notable drop in performance. These findings highlight that our training recipe is both data- and parameter-efficient, achieving strong results with minimal resource expenditure.

Training Settings		Infoseek		
		Unseen-Q	Unseen-E	All
Original		32.8	28.5	30.7
Data	4x	34.8	28.4	31.3
	2x	32.2	29.8	30.9
	0.5x	26.5	24.4	25.4
	0.25x	17.8	17.5	17.6
Parameters	4x	26.4	22.1	24.1
	2x	28.6	24.8	26.3
	0.5x	27.8	24.7	26.1
	0.25x	23.1	20.3	21.6

Table 6: Performance of CoMEM on Qwen2.5-VL under different training data and parameter settings.

628 E Case Study

629 In this appendix, we present a qualitative case study to demonstrate the effectiveness of our proposed
630 model. Given a question and a corresponding query image, our pipeline first retrieves the top 10
631 relevant image-text pairs from the WIT knowledge base to provide rich contextual information. Due
632 to space constraints, we only display three representative retrieved pairs for each example in this
633 appendix. We then compare the performance of our CoMEM model against two baselines: the
634 standalone Qwen2.5-VL and a baseline retrieval-augmented generation (RAG) model. CoMEM can
635 effectively capture key information from retrieved supporting texts, even when the exact answer is
636 not explicitly provided, and perform reasoning to derive the correct answer.









Question	Retrieved Information	Answer
<p>Q: Whom was this building officially opened by?</p> 	<p>1. Rogers Stirk Harbour + Partners: Esta lista contiene los proyectos del estudio desde su fundación en 1977 hasta la actualidad. Para los trabajos previos de Richard Rogers, Team 4, Richard and Su Rogers y Piano + Rogers, véase el artículo Richard Rogers.The Richard Rogers Partnership</p> <p>Edificio Lloyd's, Londres, Reino Unido (1978-1984)</p> <p>Fábrica Fleetguard, Quimper, Francia (1979-1981) ...</p> <p>2. Richard Rogers: Ričardas Džordžas Rodžersas (angl. Richard George Rogers, g. 1933 liepos 23 d. Florencijoje) – pasaulinio garso britų architektas. 1938 m. su tėvais persikėlė į Didžiąją Britaniją, mokėsi Architektūros Asociacijos mokykloje ...</p> <p>3. Lloyd's di Londra: L'edificio dei Lloyd's fu eretto trail 1978 e il 1986 e consiste in una forma rettangolare e in una maestosa torre di quattordici piani con una struttura in cemento, la cui misura corrisponde a 45 metri di larghezza e 67 metri di lunghezza...</p>   	<p>Qwen2.5: Nigel Lawson ✗</p> <p>Qwen2.5+RAG: No information available ✗</p> <p>Qwen2.5+CoMEM: Reasoning Process: The building in the image is the Lloyd's Building, which was officially opened by Queen Elizabeth II on 16th October 1986. Final answer: \boxed{Queen Elizabeth II} ✓</p>
Question	Retrieved Information	Answer
<p>Q: what was the date this aircraft enter into service?</p> 	<p>1. Aer Lingus: On 28 April 1958, Aerlínte Éireann operated its first transatlantic service from Shannon to New York ... The Boeing 720s proved to be a success for the airline on the transatlantic routes. To supplement these, Aer Lingus took delivery of its first larger Boeing 707 ...</p> <p>2. Saab 90 Scandia: Saab 90 Scandia oli reisilennuk, mida tootis Saab. Kuna 1944. aastal hakkas selgeks saama, et sõjategevus (Teine maailmasõda) Euroopas hakkab lõppema, siis leidis Saab, kes oli seni põhiliselt tootnud militaarvarustust, et peab oma tootevalikut laiendama. Juhatus otsustas toota ...</p> <p>3. History of United Airlines: United sought overseas routes in the 1960s, but the Transpacific Route Case (1969) denied them this expansion; it did not gain an overseas route until 1983, when they began flights to Tokyo from Portland and Seattle...</p>   	<p>Qwen2.5: 1959 ✗</p> <p>Qwen2.5+RAG: No information available ✗</p> <p>Qwen2.5+CoMEM: Reasoning Process: The aircraft in the image is a Boeing 707. The Boeing 707 entered into service on February 26, 1958. Final answer: \boxed{1958} ✓</p>

Figure 4: Case studies comparing CoMEM with baseline model and model with RAG.

637 These case studies demonstrate that CoMEM is able to generate accurate answers in challenging
638 scenarios where baseline models either fail or return incomplete information. This underscores the
639 strength of our CoMEM approach in encoding and leveraging complex multimodal and multilingual
640 knowledge, thereby enabling more robust performance in advanced reasoning tasks.

References

- [1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [2] 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.
- [3] Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Hannah R Sheahan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. Language models show human-like content effects on reasoning tasks. *arXiv preprint arXiv:2207.07051*, 2022.
- [4] Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*, 2024.
- [5] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- [6] Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. A survey on large language models for code generation. *arXiv preprint arXiv:2406.00515*, 2024.
- [7] Zichen Wen, Yifeng Gao, Weijia Li, Conghui He, and Linfeng Zhang. Token pruning in multimodal large language models: Are we solving the right problem? *arXiv preprint arXiv:2502.11501*, 2025.
- [8] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Angela Fan, Vishrav Chaudhary, Matthias Gallé, Veselin Stoyanov, and Wen-tau Yih. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [9] Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. Augmenting language models with long-term memory. *arXiv preprint arXiv:2306.07174*, 2023.
- [10] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-augmented language model pre-training. In *International Conference on Machine Learning (ICML)*, 2020.
- [11] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibor Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [12] Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*, 2024.
- [13] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pages 19–35. Springer, 2024.
- [14] Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*, 2024.
- [15] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. <https://transformer-circuits.pub/2021/framework/index.html>.

- [16] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.
- [17] Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.
- [18] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021.
- [19] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.
- [20] David Grangier and Dan Iter. The trade-offs of domain adaptation for neural language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3802–3813, Dublin, Ireland, May 2022. Association for Computational Linguistics.
- [21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.
- [22] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does bert look at? an analysis of bert’s attention. *arXiv preprint arXiv:1906.04341*, 2019.
- [23] Edward J Hu, Yelong Shen, Phil Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- [24] 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*, 2023.
- [25] Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can pre-trained vision and language models answer visual information-seeking questions? *arXiv preprint arXiv:2302.11713*, 2023.
- [26] 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*, pages 3195–3204, 2019.
- [27] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer, 2022.
- [28] Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Wiki-LLaVA: Hierarchical retrieval-augmented generation for multimodal llms. *arXiv preprint arXiv:2404.15406*, 2024.
- [29] Jingyuan Qi, Zhiyang Xu, Rulin Shao, Yang Chen, Jin Di, Yu Cheng, Qifan Wang, and Lifu Huang. Rora-vlm: Robust retrieval augmentation for vision language models. *arXiv preprint arXiv:2410.08876*, 2024.

- [30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.
- [31] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2443–2449, 2021.
- [32] Jingyuan Qi, Zhiyang Xu, Rulin Shao, Yang Chen, Jin Di, Yu Cheng, Qifan Wang, and Lifu Huang. RORA-VLM: Robust retrieval augmentation for vision language models. *arXiv preprint arXiv:2410.08876*, 2024.
- [33] Thomas Mensink, Jasper Uijlings, Lluís Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araujo, and Vittorio Ferrari. Encyclopedic vqa: Visual questions about detailed properties of fine-grained categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3113–3124, 2023.
- [34] Yibin Yan and Weidi Xie. Echosight: Advancing visual-language models with wiki knowledge. *arXiv preprint arXiv:2407.12735*, 2024. Accepted at EMNLP 2024 Findings.
- [35] Federico Cocchi, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Augmenting multimodal llms with self-reflective tokens for knowledge-based visual question answering. *arXiv preprint arXiv:2411.16863*, 2024. Accepted at CVPR 2025.
- [36] Xin Cheng, Xun Wang, Xingxing Zhang, Tao Ge, Si-Qing Chen, Furu Wei, Huishuai Zhang, and Dongyan Zhao. xrag: Extreme context compression for retrieval-augmented generation with one token. *arXiv preprint arXiv:2405.13792*, 2024.
- [37] Xubing Ye, Yukang Gan, Xiaoke Huang, Yixiao Ge, and Yansong Tang. Voco-llama: Towards vision compression with large language models. *arXiv preprint arXiv:2406.12275*, 2024.
- [38] Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. MA-LMM: Memory-augmented large multimodal model for long-term video understanding. *arXiv preprint arXiv:2404.05726*, 2024.
- [39] Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12065–12075, 2023.
- [40] Wenbo Hu, Jia-Chen Gu, Zi-Yi Dou, Mohsen Fayyaz, Pan Lu, Kai-Wei Chang, and Nanyun Peng. Mrag-bench: Vision-centric evaluation for retrieval-augmented multimodal models. *arXiv preprint arXiv:2410.08182*, 2024.
- [41] Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, José G Moreno, and Jesús Lovón Melgarejo. Viquae, a dataset for knowledge-based visual question answering about named entities. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 3108–3120, 2022.
- [42] David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, et al. CVQA: Culturally-diverse multilingual visual question answering benchmark. *arXiv preprint arXiv:2406.05967*, 2024.
- [43] OpenAI. Gpt-4o technical report, 2024. Accessed: 2025-05-10.
- [44] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.
- [45] LLaVA-VL Team. Llava-next: Open large multimodal models. <https://github.com/LLaVA-VL/LLaVA-NeXT>, 2024. Accessed: 2025-05-10.

- [46] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, Songyang Zhang, Wenwei Zhang, Yining Li, Yang Gao, Peng Sun, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, Hang Yan, Conghui He, Xingcheng Zhang, Kai Chen, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024.
- [47] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*, 2024.
- [48] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.
- [49] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- [50] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.
- [51] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruvi Shah, Xianzhi Du, Futang Peng, Anton Belyi, et al. Mm1: methods, analysis and insights from multimodal llm pre-training. In *European Conference on Computer Vision*, pages 304–323. Springer, 2024.
- [52] Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, et al. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27091–27101, 2024.
- [53] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.
- [54] Jesse Mu, Xiang Li, and Noah Goodman. Learning to compress prompts with gist tokens. *Advances in Neural Information Processing Systems*, 36:19327–19352, 2023.
- [55] Xiangfeng Wang, Zaiyi Chen, Zheyong Xie, Tong Xu, Yongyi He, and Enhong Chen. In-context former: Lightning-fast compressing context for large language model. *arXiv preprint arXiv:2406.13618*, 2024.
- [56] Cangqing Wang, Yutian Yang, Ruisi Li, Dan Sun, Ruicong Cai, Yuzhu Zhang, and Chengqian Fu. Adapting llms for efficient context processing through soft prompt compression. In *Proceedings of the International Conference on Modeling, Natural Language Processing and Machine Learning*, pages 91–97, 2024.
- [57] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.