You Never Know: Quantization Induces Inconsistent Biases in Vision-Language Foundation Models

Eric Slyman, Anirudh Kanneganti, Sanghyun Hong, Stefan Lee
Oregon State University
{slymane, kannegaa, hongsa, leestef}@oregonstate.edu

Abstract

We study the impact of a standard practice in compressing foundation vision-language models–*quantization*–on the models' ability to produce socially-fair outputs. In contrast to prior findings with unimodal models that compression consistently amplifies social biases, our extensive evaluation of four quantization settings across three datasets and three CLIP variants yields a surprising result: while individual models demonstrate bias, we find *no consistent change* in bias magnitude or direction across a population of compressed models due to quantization.

1 Introduction

Quantization (Gholami et al., 2022) is a leading practice in compressing deep learning models: it transforms a model's parameter representation from 32-bit floating-point numbers to lower bit-width (e.g., 8-bit or 4-bit integers), thereby reducing memory footprint and inference latency significantly. However, these transformations in number representation can introduce small numerical perturbations to a model's parameter values, potentially leading to undesirable behaviors of a model after quantization (Hong et al., 2024; Tian et al., 2022; Hooker et al., 2019). In this paper, we study the effects of quantization on the fairness outcomes of foundation vision-language (ViL) models.

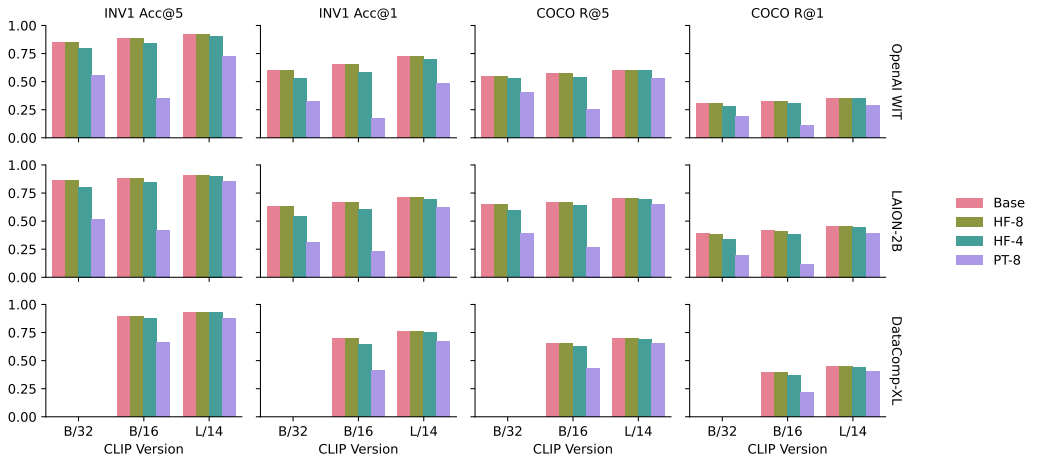
Related work. Most work studies compression-induced bias in the unimodal setting, such as vision or language models. Hooker et al. (2019) first noted that the drop in accuracy induced by compressing vision models is concentrated in a few classes which are "*cannibalized*" to preserve accuracy in the others. Follow-up work (Hooker et al., 2020) notes that compression error disparately affects data with low representation in the training distribution that often correlates with socially meaningful features like gender and age. Silva et al. (2021) similarly find that distilled language models "*almost always exhibit statistically significant bias.*" Subsequent works show that compressing language models amplifies gender bias (Renduchintala et al., 2021; Ahn et al., 2022) and that vision model compression has a disparate impact on face classification accuracy (Tran et al., 2022), expression recognition (Stoychev and Gunes, 2022), and other traditional vision tasks (Paganini, 2023).

Recent studies extend these investigations to varied compression techniques across different domains such as facial recognition (Lin et al., 2022), medical diagnosis (Wu et al., 2022), and multilingual NLP (Ogueji et al., 2022; Lee and Lee, 2023). Fairness-aware compression methods analyzed the trade-offs between model fairness, performance, and environmental impact (Hessenthaler et al., 2022). Yi-Lin Sung (2024) have even developed a compression technique specifically for ViL models. However, no work to date has studied the fairness impacts of compression for multimodal ViL models, leaving a critical gap on how these techniques affect integrated architectures.

Contributions. We address this knowledge gap by extensively evaluating quantization effects in multimodal ViL models focusing on fairness outcomes across socially meaningful features like gender, age, and race. Contrary to previous findings in fair compression, our analysis reveales a sur-

Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models
38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Figure 1: Zero-shot image classification accuracy on ImageNet1K (Deng et al., 2009) and text-image retrieval recall on COCO Captions Lin et al. (2014) across varied CLIP versions, training data sources, and quantization methods. Higher (↑) is better in all cases. HuggingFace-based quantization methods preserve performance while the PyTorch-based method shows a reduction across metrics.



prising result: there is no significant evidence for consistent bias amplification across quantized ViL models. While individual models do exhibit biases, the direction and magnitude of these biases were not uniform, suggesting that the impact of compression on fairness may be more nuanced in multimodal contexts. These findings raise questions about the generalizability of compression-induced bias across different model architectures and modalities, indicating a potential need for a more refined understanding of how quantization affects fairness. Additionally, this result may indirectly support recent findings on arbitrariness in fair binary classification (Cooper et al., 2024).

2 Methodology and Experiments

We examine three common off-the-shelf quantization methods for model compression: 8-bit and 4-bit quantization from the `bitsandbytes` (Dettmers et al., 2022b) integration into Hugging Face Transformers (Wolf et al., 2020), and PyTorch's (Paszke et al., 2019) 8-bit dynamic quantization.

HuggingFace 8-bit Quantization. The 8-bit quantization method, initially introduced by Dettmers et al. (2022a) in their work on LLM.int8(), represents a significant step towards efficient model compression. This method employs a linear quantization scheme for weight representation, quantizing weights to 8-bit integers while preserving activations in higher precision. It offers up to 50% reduction in model size compared to FP16 representations and generally yields better performance than lower-bit alternatives, albeit with a smaller compression ratio. A key innovation in this approach is the use of vector-wise quantization, which quantizes vectors (rows or columns) of the weight matrix independently, allowing for better preservation of the weight distribution.

HuggingFace 4-bit Quantization. Building upon this work, Dettmers et al. (2023) introduced 4-bit quantization with their QLoRA method, which utilizes the NormalFloat (NF4) data type. This specialized format is optimized for the weight distribution typically observed in neural networks. The 4-bit quantization approach achieves up to 75% reduction in model size compared to FP16 representations, enabling the loading and inference of larger models on consumer-grade GPUs with limited memory. A crucial aspect of this method is the use of blockwise quantization. In this scheme, the weight matrix is divided into small blocks (typically 64 or 128 elements), and each block is quantized independently. This approach allows for more fine-grained quantization, better preserving the local structure of the weight matrix.

Pytorch 8-bit Quantization. PyTorch's dynamic quantization (Paszke et al., 2019) offers a complementary approach that's worth considering. This post-training technique focuses on reducing inference time and memory usage, particularly on CPU architectures. It quantizes weights to 8-bit integers and dynamically quantizes activations during the inference phase, utilizing a dynamic range for activations by calculating scaling factors on-the-fly. This method is well-suited for models with varying input sizes or dynamic computation graphs.

Table 1: P-value and 95% confidence intervals for paired t-tests on FACET disparity pre- and post-quantization difference in means. Cells in yellow/green are significant at 90%/95% confidence.

Quant Method	Demographic	Min Disp@1		Max Disp@1		Min Disp@5		Max Disp@5	
		p	$95\%CI$	p	$95\%CI$	p	$95\%CI$	p	$95\%CI$
HuggingFace	Gender (MF)	.862	-.01, .01	.082	-.02, .00	.246	-.01, .00	.356	-.05, .11
4-Bit	Skin Tone (LD)	.307	-.02, .04	.035	.01, .10	.164	-.01, .03	.210	-.04, .12
	Age (MY)	.774	-.01, .01	.425	-.04, .08	.133	-.01, .00	.587	-.05, .09
	Age (MO)	.262	-.01, .00	.141	-.13, .03	.501	.00, .00	.388	-.04, .09
HuggingFace	Gender (MF)	.310	.00, .01	.945	-.01, .01	.999	.00, .00	.502	-.03, .05
8-Bit	Skin Tone (LD)	.700	-.01, .02	.843	-.05, .06	.817	-.01, .01	.447	-.05, .03
	Age (MY)	.765	.00, .00	.868	-.03, .03	.456	.00, .00	.245	-.01, .02
	Age (MO)	.396	-.01, .00	.269	-.03, .01	.340	.00, .01	.444	-.06, .12
PyTorch	Gender (MF)	.115	-.01, .03	.192	-.05, .16	.278	-.03, .01	.076	-.02, .20
8-Bit	Skin Tone (LD)	.205	-.01, .04	.247	-.07, .17	.399	-.04, .02	.158	-.08, .02
	Age (MY)	.117	-.01, .03	.850	-.17, .15	.095	-.02, .00	.940	-.23, .22
	Age (MO)	.045	.00, .04	.462	-.14, .24	.159	-.02, .01	.348	-.26, .13

We choose CLIP (Radford et al., 2021) as a representative foundation alignment model and apply each quantization method above to a variety of model variants across different training data sources. Specifically, we consider the base size vision transformer (ViT) (Dosovitskiy, 2020) variants with sequence length 32 (B/32) and 16 (B/16), and the large size ViT variant with sequence length 14 (L/14). For each CLIP variant, we consider a model pretrained on OpenAI WIT (Radford et al., 2021), LAION-2B (Schuhmann et al., 2022), and DataComp-XL (Gadre et al., 2024). We find the weights available online for the B/32 varient trained on DataComp-XL to be corrupted, resulting in an evaluation over eight distinct models with three quantization methods totalling 32 scenarios.

2.1 Evaluation Datasets and Metrics

We evaluate across three benchmarks to validate if quantized models are both *accurate* and *fair*.

Zero-Shot Classification and Retrieval. We evaluate the accuracy of each model and its quantized variants across two common benchmark tasks for foundation vision-language alignment models: zero-shot image classification on ImageNet (Deng et al., 2009) and text-based image retrieval on COCO (Lin et al., 2014). Quantized variants *should* have similar accuracy to the original model.

Fair Zero-Shot Classification. The FACET (Gustafson et al., 2023) dataset contains expert image annotations for 52 person related classes, age, skin tone, and gender presentation. We perform zero-shot classification over the person-classes by constructing a text prompt for each class and predicting the class used to construct the prompt with highest similarity to the image. Following Gustafson et al. (2023), we measure the disparity between pairs of values within a sensitive group (*e.g.* $\{light,\ dark\} \in skintone$) as the difference in recall between true-positive instances of a person-class for each value. A large magnitude disparity indicates that a model better predicts positive instances for one member of the group, while a disparity of zero indicates group *equality of opportunity*. We study the the maximum and minimum disparity measured across person-classes. We utilize the same sensitive groups as Slyman et al. (2024), evaluating perceived gender expression by masculine *vs*. feminine presentation, lighter (1-4MST[1]) *vs*. darker (6-10MST) skin tone, and middle *vs*. younger/older age for all person-classes which have at least 25 samples in both subgroups.

Fair Image Retrieval. FairFace (Kärkkäinen and Joo, 2021) annotates cropped faces with perceived race, age, and gender. Following (Seth et al., 2023), we assess how much the top-k results of an image-text query differ across sensitive attribute values in the validation set using MaxSkew@k (Geyik et al., 2019). For a given top-k image set τ_r^k from query r, let $P_{\tau_r^k,a_i} \in [0,1]$ be the actual proportion of images with a particular value $a_i \in A$ of sensitive attribute A, and $P_{r,a_i} \in [0,1]$

[1]Monk Skin Tone scale Monk (2019)

be the desired proportion estimated from true rates in the full dataset. Then the skew for a_i is:

$$Skew_{a_i}@k(\tau_r) = \ln\left(\frac{P_{\tau_r^k,a_i}}{P_{r,a_i}}\right) \quad (1)$$

Skew@k is specific to a single value of a sensitive attribute. To provide a more comprehensive view, we report the most (least) skewed attribute value as MaxSkew@k (MinSkew@k). MaxSkew@k indicates the "*largest unfair advantage*" (Geyik et al., 2019) given to images with a particular attribute value in the top-k results while MinSkew@k captures the "*worst disadvantage in representation*" for a subgroup. With the condition that the desired proportion of images matches the true distribution of those images in the dataset, an optimal MaxSkew@k of 0 can be shown to achieve the fairness criteria of *demographic parity*. Following Berg et al. (2022), we report average MaxSkew@1000 across 240 (un)favorable captions orthogonal to images in the dataset, matching test attributes and prompts for race, age, and gender. Similar to Slyman et al. (2024), we reduce noise by binning age into: *younger* (0-19), *middle* (20-49), and *older* (50-70+) subgroups.

Table 2: P-value and confidence interval for paired t-tests on FairFace skew pre- and post-quantization means. Cells in yellow/green are significant at 90%/95% confidence.

Quant Method	**Demographic**	**MinSkew**			**MaxSkew**		
		p	95%CI		p	@95%CI	
HuggingFace	Gender	.066	-.19,	.01	.039	-.07,	.01
4-Bit	Race	.225	-.04,	.12	.073	-.01,	.06
	Age	.557	-.10,	.06	.800	-.04,	.03
HuggingFace	Gender	.632	-.01,	.02	.655	-.01,	.01
8-Bit	Race	.717	-.01,	.01	.060	.00,	.01
	Age	.153	-.02,	.01	.157	-.01,	.01
PyTorch	Gender	.560	-.08,	.13	.535	-.03,	.05
8-Bit	Race	.938	-.13,	.12	.555	-.03,	.06
	Age	.077	-.02,	.26	.124	-.01,	.07

3 Empirical Evaluation

Accuracy. As shown in Fig. 1, the selected quantization methods generally preserve accuracy across different models and tasks. This result indicates that the methods chosen for our study are effective in terms of preserving baseline performance. A method which does not preserve performance (*e.g.* resulting in random predictions) may otherwise trivially fulfil many common fairness criteria.

Fairness. We consider fairness evaluations as paired pre-post/quantization measurements and assess the significance of difference between the two measures with a paired t-test. The results for fairness are mixed. Table 1 presents the outcomes for FACET, where we observe inconsistent equality of opportunity outcomes in zero-shot image classification. The disparities across different demographic groups vary, with some quantization methods leading to minor but somewhat statistically significant changes. As shown in Table 2, FairFace demonstrates similar inconsistent skew outcomes in image retrieval. We note that these observations are without correcting for multiple testing, and that the minor significant results observed here are likely to disappear under most correction methods.

Limitations. Our evaluations make several generalizability limiting assumptions. Specifically, we study only CLIP models under quantization as the compression method. It would be a compelling line of future work to understand fairness outcomes when applying more advanced compression methods (*e.g.*, pruning or distillation), alternative alignment model architectures, or more advanced ViL models (*e.g.* BLIP (Li et al., 2023)) which can perform VQA and image captioning tasks.

4 Conclusion

Our study reveals that the impact of quantization on bias in multimodal ViL models is neither consistent nor uniform across different models, methods, and datasets. The direction and magnitude of bias introduced by quantization varied, indicating that its effects on fairness are complex and context-dependent. These findings challenge the assumption that quantization consistently influences bias across settings, highlighting the need for a more nuanced understanding of how compression techniques impact fairness across diverse model architectures and applications.

Acknowledgments. Anirudh and Sanghyun are supported by the Samsung Global Research Outreach (GRO) Program 2024. Any opinions, findings, and conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of the funding agencies.

References

Jaimeen Ahn, Hwaran Lee, Jinhwa Kim, and Alice Oh. Why knowledge distillation amplifies gender bias and how to mitigate - from the perspective of distilbert. *GeBNLP*, 2022. 1

Hugo Berg, Siobhan Hall, Yash Bhalgat, Hannah Kirk, Aleksandar Shtedritski, and Max Bain. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 806–822, 2022. 4

A Feder Cooper, Katherine Lee, Madiha Choksi, Solon Barocas, Christopher De Sa, James Grimmelmann, Jon Kleinberg, Siddhartha Sen, and Baobao Zhang. Arbitrariness and prediction: The confounding role of variance in fair classification. *AAAI*, 2024. 2

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2, 3

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Gpt3.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, pages 30318–30332, 2022a. 2

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 8-bit optimizers via block-wise quantization. In *International Conference on Learning Representations*, 2022b. 2

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, pages 10088–10115, 2023. 2

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024. 3

Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, page 2221–2231, 2019. 3, 4

Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W Mahoney, and Kurt Keutzer. A survey of quantization methods for efficient neural network inference. In *Low-Power Computer Vision*, pages 291–326. Chapman and Hall/CRC, 2022. 1

Laura Gustafson, Chloe Rolland, Nikhila Ravi, Quentin Duval, Aaron Adcock, Cheng-Yang Fu, Melissa Hall, and Candace Ross. Facet: Fairness in computer vision evaluation benchmark. In *International Conference on Computer Vision*, pages 20370–20382, 2023. 3

Marius Hessenthaler, Emma Strubell, Dirk Hovy, , and Anne Lauscher. Bridging fairness and environmental sustainability in nlp. *EMNLP*, 2022. 1

Sanghyun Hong, Michael Panaitescu, Yiğitcan Kaya, and Tudor Dumitraş. Qu-anti-zation: exploiting quantization artifacts for achieving adversarial outcomes. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2024. Curran Associates Inc. 1

Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. What do compressed deep neural networks forget? *arXiv*, 2019. 1

Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. Characterising bias in compressed models. *arXiv*, 2020. 1

Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021. 3

Jiwoon Lee and Jaeho Lee. Debiased distillation by transplanting the last layer. *arXiv*, 2023. 1

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. 4

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 *European Conference on Computer Vision*, pages 740–755, 2014. 2, 3

Xiaofeng Lin, Seungbae Kim, and Jungseock Joo. Fairgrape: Fairness-aware gradient pruning method for face attribute classification. *ECCV*, 2022. 1

Ellis Monk. Monk skin tone scale, 2019. 3

Kelechi Ogueji, Orevaoghene Ahia, Gbemileke Onilude, Sebastian Gehrmann, Sara Hooker, , and Julia Kreutzer. Intriguing properties of compression on multilingual models. *EMNLP*, 2022. 1

Michela Paganini. Prune responsibly. *arXiv*, 2023. 1

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 2019. 2

Alec Radford, Jong Wook Kim, Chris Hallacy, A. 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 *International Conference on Machine Learning*, 2021. 3

Adithya Renduchintala, Denise Diaz, Kenneth Heafield, Xian Li, and Mona Diab. Gender bias amplification during speed-quality optimization in neural machine translation. *IJCNLP*, 2021. 1

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Neural Information Processing Systems*, 35:25278–25294, 2022. 3

Ashish Seth, Mayur Hemani, and Chirag Agarwal. Dear: Debiasing vision-language models with additive residuals. In *Computer Vision and Pattern Recognition*, pages 6820–6829, 2023. 3

Andrew Silva, Pradyumna Tambwekar, , and Matthew Gombolay. Towards a comprehensive understanding and accurate evaluation of societal biases in pre-trained transformers. *NAACL*, 2021. 1

Eric Slyman, Stefan Lee, Scott Cohen, and Kushal Kafle. Fairdedup: Detecting and mitigating vision-language fairness disparities in semantic dataset deduplication. *Computer Vision and Patern Recognition*, 2024. 3, 4

Samuil Stoychev and Hatice Gunes. The effect of model compression on fairness in facial expression recognition. *ICPR*, 2022. 1

Yulong Tian, Fnu Suya, Fengyuan Xu, and David Evans. Stealthy backdoors as compression artifacts. *IEEE Transactions on Information Forensics and Security*, 17:1372–1387, 2022. 1

Cuong Tran, Ferdinando Fioretto, Jung-Eun Kim, and Rakshit Naidu. Pruning has a disparate impact on model accuracy. *NeurIPS*, 2022. 1

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020. 2

Yawen Wu, Dewen Zeng, Xiaowei Xu, Yiyu Shi, and Jingtong Hu. Fairprune: Achieving fairness through pruning for dermatological disease diagnosis. *ECCV*, 2022. 1

Mohit Bansal Yi-Lin Sung, Jaehong Yoon. Ecoflap: Efficient coarse-to-fine layer-wise pruning for vision-language models. In *International Conference on Learning Representations (ICLR)*, 2024. 1