Two Effects, One Trigger: On the Modality Gap, Object Bias, and Information Imbalance in Contrastive Vision-Language Representation Learning

Simon Schrodi[1]*, David T. Hoffmann[1,2]*, Max Argus[1], Volker Fischer[2] & Thomas Brox[1]
[1] University of Freiburg, [2] Bosch Center for Artificial Intelligence

Abstract

Contrastive vision-language models like CLIP have gained popularity for their rich representations, that are applicable in various downstream tasks. Despite their successes in some tasks, like zero-shot image recognition, they perform surprisingly poor on other tasks, like attribute detection. Previous work has attributed these challenges to the modality gap, a separation of image and text in the shared representation space, and a bias favoring objects over other factors, such as attributes. We investigate both phenomena. Specifically, we find an unintuitive correlation between the modality gap and downstream performance, with only a few embedding dimensions driving the gap. But how to determine what leads to the emergence of these phenomena? To answer this question we design a controlled setting which allows us to control the amount of shared information between the modalities. This revealed that the driving factor behind both, the modality gap and the object bias, is the information imbalance between images and captions.

1 Introduction

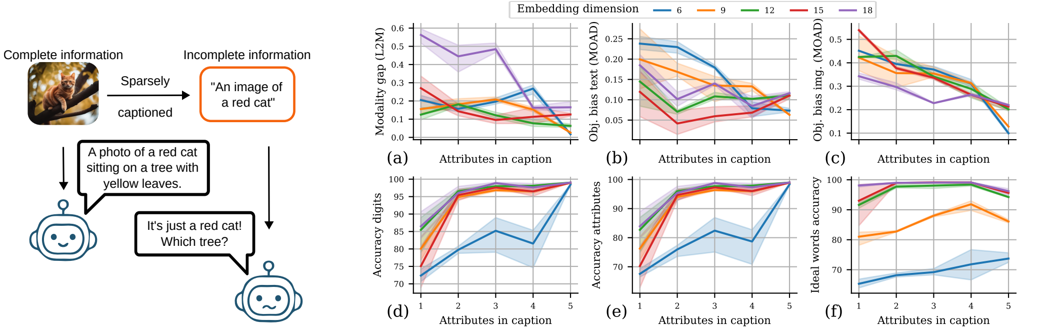
Large-scale Vision-Language Models (VLMs) have become increasingly popular and are successfully applied to numerous tasks. Their great advantage lies in the ability to exploit weak supervision, which can be obtained with low costs by scraping the internet for image-text pairs. VLMs are commonly trained with a contrastive loss (Radford et al., 2021; Jia et al., 2021). Despite the successes of such multi-modal models, recent works unveiled various undesired characteristics: modality gap in the joint embedding space (Liang et al., 2022) or a bias towards objects at the cost of other factors like attributes (So et al., 2023; Zhou et al., 2023; Bravo et al., 2023). But how bad are these effects? To date, both the consequences and the underlying triggers are not fully understood.

In this work, we surprisingly find that *larger modality gap correlates with better performance* in a large-scale study, only very *few embedding dimensions drive the modality gap*, and the embeddings of the modalities have *distinct characteristics*. Next, we confirm that contrastive VLMs are more *biased towards objects* than attributes. However, word frequency is not the cause and we link it to the presence of words in captions - a *caption presence bias* taking a per-sample view. We find that the caption presence bias stems from an *information imbalance* between modalities: the image modality commonly contains more information than the text, while the text modality determines the focal point. Refer to Fig. 1a for an illustration. Finally, we validate this explanation in a synthetic setting, in which we control the data-generating process. Our findings provide practical utility as well as justification to enrich visual captions for contrastive training of VLMs.

2 Analysis of the modality gap

Recent work by Liang et al. (2022) revealed the existence of a *modality gap* within the shared representation space of multi-modal models, i.e., the embeddings of the modalities are located in two completely separate regions. They defined the modality gap distance as the L2-distance between

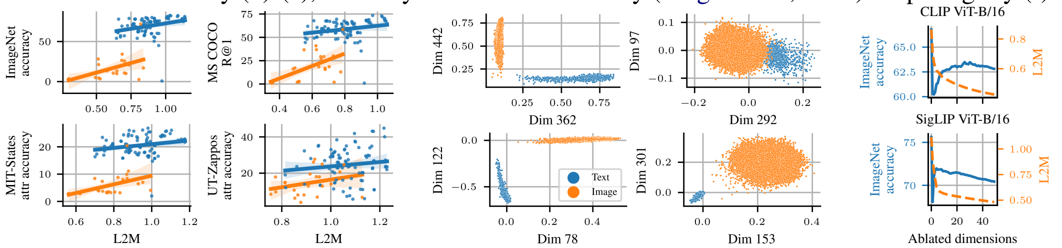
[*] Equal contribution. Correspondence to: {schrodi,hoffmann}@cs.uni-freiburg.de.



(a) Sketch for information imbalance between the modalities.

(b) Increasing shared information between modality inputs improves representations.

Figure 1: a: Information imbalance makes it impossible for the image encoder (left) to know what a (sparse) caption may contain. Thus, it may focus on the most salient objects due to their high probability of being present in the caption and may tend to neglect other more unlikely factors, such as attributes. b: We trained multiple small CLIP models on our dataset MAD (see Appendix C). The number of attributes present in the captions controls the information imbalance between the modalities. Less information imbalance leads to a smaller modality gap (a), object bias (b)-(c), higher downstream accuracy (d)-(e), and only ideal words accuracy (Trager et al., 2023) drops slightly (f).



(a) Relation between modality gap and performance.

(b) Some pairs of embedding dimensions can perfectly separate images from text.

(c) Successive ablation of dimensions.

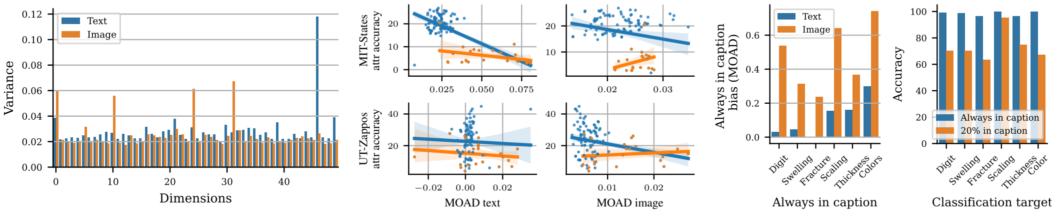
Figure 2: a: plot of modality gap (L2M) vs. performance for 113 contrastive VLMs. To factor out the influence of the training dataset size, we split the models in two groups, i.e., medium- and large-scale. b: two embedding dimensions suffice to perfectly separate image from text. Plotted for ImageNet data. c: Successive removal of embedding dimensions based on maximal reduction in L2M leads to a sharp drop, followed by a partial recovery of ImageNet accuracy.

the Means (L2M) of the embeddings: $||\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i - \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i||$, where $\mathbf{x}_i$ and $\mathbf{y}_i$ are the i-th L2-normalized image or text embeddings, respectively.

With increasing modality gap comes improvements in performance. We evaluated the modality gap distance L2M and downstream task performance (DTP) for a total of 113 VLMs trained with contrastive loss. Refer to Appendix B for details on the models and downstream tasks. While theoretical work (Wang & Isola, 2020) suggests that there may be no modality gap in the asymptotic limit with infinite negatives of the contrastive loss, we surprisingly find the contrary: the increase of the modality gap distance co-occurs with the improvement of DTP; see Fig. 2a. Moreover, we observe a clear separation of models trained on medium- (i.e., $\leq$128 M image-text pairs) and large-scale data. We conjecture that the dataset may be an important forming factor of the modality gap due to, e.g., worse image-text alignments.

Takeaway 1: The modality gap distance (according to L2M) increases as DTP improves.

Few dimensions drive the modality gap. Fig. 2b shows that few embedding dimensions alone are sufficient to separate images from texts. One may think that ablating these dimensions should close the gap and lead to performance improvements, but we observe the opposite in Fig. 2c: DTP initially drops sharply and recovers partially. We suspect that the re-normalization of the ablated embeddings causes substantial changes in cosine similarities and cross-modal neighborhoods. Indeed, the most modality-separating dimensions are characterized by large values in one modality and small ones in the other (see next section) that may cause such substantial changes.



(a) Absolute mean of CLIP ViT-B/16 for 50 embedding dimensions.

(b) Relation between object bias and downstream task performance.

(c) Connecting caption presence bias with object bias and DTP.

Figure 3: a: Most embedding dimensions have similar mean values for both modalities but some have substantial differences. We find that these correspond to the most modality-separating dimensions. Image-text samples from MS COCO. b: DTP vs. object bias for a total of 113 contrastive VLMs. We find a bias towards objects (mostly positive MOAD values) but only a weak correlation with DTP. c: When showing a particular word (object or attribute) in the caption during the training, the model is biased towards it (left) and achieves higher accuracy for it (right). Thus, the "object bias" is just a caption presence bias. Note that the bias is larger for the image encoder, as it needs to match to the most likely caption, as sketched in Fig. 1a.

Takeaway 2: Few dimensions drive the modality gap and can separate the modalities.

3 Differences in the embeddings

In this section we reveal various differences of image and text embeddings, particularly the different organization thereof; even though one may suspect that both embeddings are similarly organized (Trager et al., 2023). We find that the mean of the absolute mean embedding is similar for images and texts, i.e., 0.0282 and 0.0267, respectively. However, Fig. 3a reveals while most embedding dimensions have comparable absolute means, there are few notable exceptions. Even more, we find that these dimensions (typically) correspond to the most-modality separating dimensions (Fig. 2b). We suspect that these are an result of the contrastive loss optimization with image-text misalignments, i.e., the uniformity term may be optimized by making the modalities as dissimilar as possible as the alignment term's maximization is limited by misalignments.

Table 1: **Dissimilarity of neighborhood orderings in the embedding space.** Kendall-Tau distance (KTD) $\in [0, 1]$ lower is better. C: Cifar, IN: ImageNet, s.:split.

Dataset	KTD (↓)
C-10	0.3399
C-100	0.4965
IN-100 s. 1	0.4975
IN-100 s. 2	0.5046
IN-100 s. 3	0.5081

Following the ideal words approach of Trager et al. (2023) and extending it to ideal images on MIT-States & UT-Zappos, we find low cosine-simiarities (0.19 & 0.16) between the ideal images and ideal words. However, when we correct them with the modality gap vector (mean difference vector between matching image and text embeddings) cosine similarities significantly increase (0.56 & 0.40). Hence, this indicates that embedding directions of each modality have different "meanings" when not corrected by the modality gap vector. Further, we tested similarity of neighborhood relations in the embedding spaces by computing the normalized Kendall-Tau distance on the mean vectors for each class from CIFAR-10, CIFAR-100 (Krizhevsky et al., 2009), and the three ImageNet-100 splits from Hoffmann et al. (2022). We find that neighborhood orderings are not preserved across the modalities' embeddings (Table 1).

Takeaway 3: Direction of image and text embeddings align when corrected by the modality gap vector and neighborhood relations vary between the modalities.

4 Is object bias a mere caption presence bias?

While previous work mostly assessed object bias based on DTP benchmark results, we propose a novel similarity measure Matching Object Attribute Distance (MOAD) to explicitly quantify the object bias: $\frac{1}{4n}\sum_{i=1}^{n}\left(\mathbf{x}_{i,o}^T\mathbf{y}_{i,o} - \frac{1}{n-1}\sum_{j=1,j\neq i}^{n}\mathbf{x}_{i,o}^T\mathbf{y}_{j,o}\right) - \frac{1}{4n}\sum_{i=1}^{n}\left(\mathbf{x}_{i,a}T\mathbf{y}_{i,a} - \frac{1}{n-1}\sum_{j=1,j\neq i}^{n}\mathbf{x}_{i,a}^T\mathbf{y}_{j,a}\right)$.

In words, MOAD is the difference of alignment scores for objects o and attributes a. Negative values indicate bias towards attributes, positive values a bias towards objects, and 0 no bias. Fig. 3b affirms that contrastive VLMs are biased towards objects and models trained on large-scale data tend to have weaker object bias. However, word frequency is not the cause, as revealed by Fig. 5 in the Appendix. Interestingly, we find that improvements in object-based DTPs correlate with improved attribute detection (Fig. 6 in the Appendix).

Takeaway 4: Contrastive VLMs trained on large-scale data tend to have a lower object bias. Our results indicate that object performance improvements transfer to performance gains for other factors, such as attributes.

5 Information imbalance in image captions

We hypothesize that both modality gap and object bias are caused by an *information imbalance between image and text encoder triggered by sparse captions*. Specifically, while the image encoder has all the information about the latent factors in the image (i.e., objects, attributes, etc.), the textual descriptions are typically very sparse, i.e., only the most salient objects and perhaps a handful of attributes or similar are present in a caption. Refer to Fig. 1a for an illustration of the information imbalance problem. Consequently, the maximization of the alignment term is limited, as the image encoder cannot infer what the text encoder may encode for a given sparse captions. The best both encoders can do is to make sure that latent factors with small conditional caption presence probability contribute only little to the loss by encoding them in few dimensions, with small values or both, and focus on the most salient parts, i.e., the objects. Moreover, we posit that the image encoder should exhibit a larger caption presence bias, since the text encoder can simply encode all given information, while the image encoder needs to match to the most likely caption. We validated our hypothesis on MAD (refer to Appendix C for dataset & training details).

To confirm that the object bias is an “always in caption” bias, we modified the prevalence of the latent factors of MAD, i.e., making each digit or attribute always present while designating the others as less often present, and trained small-scale CLIP models on this data. Fig. 3c confirms this. Next, to further validate our hypothesis, we control the information imbalance in MAD. Specifically, we change the number of present attributes in the captions and ensure that the object, i.e., MNIST digit, is always present. Fig. 1b(b)-(e) show that object bias reduces while performance improves when information imbalance is reduced, as the attributes enrich the captions. Further, Fig. 1b(a) shows that the modality gap also reduces with reducing information imbalance. Visually, the reduction of the gap is also visualized in the UMAP embeddings in Fig. 7 in the Appendix. In fact, even though there exists a modality gap after model initialization, the contrastive objective within the full information setting is capable of reducing it substantially. Interestingly, we find that zero-shot accuracy using ideal words (Trager et al., 2023) decreases slightly for the full information setting (Fig. 1b(f)) but leave further investigation for future work.

Takeaway 5: Bias towards concepts, e.g., objects, is caused by their high presence probability in captions if said concept appears in the image. Reducing the information imbalance between modalities mitigates both the modality gap and object bias.

6 Conclusion

This work studied contrastive VLMs and found that surprisingly performance improves as the modality gap widens (according to L2M), the gap is driven by just a few embedding dimensions, and affirmed their object bias. We show that both phenomena are triggered by an information imbalance between modalities and a reduction of such mitigates them.

Acknowledgments. This research was funded by the Bundesministerium für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz (BMUV, German Federal Ministry for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection) based on a resolution of the German Bundestag (67KI2029A), the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under grant number 417962828, and the Bosch Center for Artificial Intelligence.

REFERENCES

Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. Evaluating CLIP: Towards Characterization of Broader Capabilities and Downstream Implications. *arXiv*, 2021.

Ibrahim Alabdulmohsin, Xiaohua Zhai, Alexander Kolesnikov, and Lucas Beyer. Getting ViT in Shape: Scaling Laws for Compute-Optimal Model Design. In *NeurIPS*, 2023.

Maria A. Bravo, Sudhanshu Mittal, Simon Ging, and Thomas Brox. Open-vocabulary Attribute Detection. In *CVPR*, 2023.

Justin Brody. On the Potential of CLIP for Compositional Logical Reasoning. In *ICLP*, 2023.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. COYO-700M: Image-Text Pair Dataset, 2022. URL https://github.com/kakaobrain/coyo-dataset.

Daniel C. Castro, Jeremy Tan, Bernhard Kainz, Ender Konukoglu, and Ben Glocker. Morpho-MNIST: Quantitative Assessment and Diagnostics for Representation Learning. *JMLR*, 2019.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts. In *CVPR*, 2021.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme Ruiz, Andreas Peter Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. PaLI: A Jointly-Scaled Multilingual Language-Image Model. In *ICLR*, 2023.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv*, 2015.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, 2023.

Guillaume Couairon, Matthijs Douze, Matthieu Cord, and Holger Schwenk. Embedding Arithmetic of Multimodal Queries for Image Retrieval. In *CVPR*, 2022.

Jonathan Crabbé, Pau Rodríguez, Vaishaal Shankar, Luca Zappella, and Arno Blaas. Robust multimodal models have outlier features and encode more concepts. *arXiv*, 2023.

Imant Daunhawer, Alice Bizeul, Emanuele Palumbo, Alexander Marx, and Julia E. Vogt. Identifiability Results for Multimodal Contrastive Learning. In *ICLR*, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 2020.

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. In *Datasets and Benchmarks Track@NeurIPS*, 2023.

Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. In *NeurIPS*, 2021.

Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal Neurons in Artificial Neural Networks. *Distill*, 2021.

Kimia Hamidieh, Haoran Zhang, Thomas Hartvigsen, and Marzyeh Ghassemi. Identifying Implicit Social Biases in Vision-Language Models. *Workshop@ICLR*, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016.

David T. Hoffmann, Nadine Behrmann, Juergen Gall, Thomas Brox, and Mehdi Noroozi. Ranking Info Noise Contrastive Estimation: Boosting Contrastive Learning via Ranked Positives. In *AAAI*, 2022.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, 2021.

Phillip Isola, Joseph J. Lim, and Edward H. Adelson. Discovering States and Transformations in Image Collections. In *CVPR*, 2015.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *ICML*, 2021.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning Multiple Layers of Features from Tiny Images, 2009.

Yann LeCun. The MNIST database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Xianhang Li, Zeyu Wang, and Cihang Xie. CLIPA-v2: Scaling CLIP Training with 81.1% Zero-shot ImageNet Accuracy within a $10,000 Budget; An Extra $4,000 Unlocks 81.8% Accuracy. *Workshop@NeurIPS*, 2023a.

Xianhang Li, Zeyu Wang, and Cihang Xie. An Inverse Scaling Law for CLIP Training. In *NeurIPS*, 2023b.

Xianhang Li, Zeyu Wang, and Cihang Xie. Grounding Visual Illusions in Language: Do Vision-Language Models Perceive Illusions Like Humans? In *EMLNP*, 2023c.

Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling Language-Image Pre-training via Masking. In *CVPR*, 2023d.

Paul Pu Liang, Zihao Deng, Martin Ma, James Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized Contrastive Learning: Going Beyond Multi-view Redundancy. In *NeurIPS*, 2023.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning. In *NeurIPS*, 2022.

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.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *CVPR*, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *ICLR*, 2019.

Joanna Materzyńska, Antonio Torralba, and David Bau. Disentangling visual and written concepts in CLIP. In *CVPR*, 2022.

Prasanna Mayilvahanan, Thaddäus Wiedemer, Evgenia Rusak, Matthias Bethge, and Wieland Brendel. Does CLIP's Generalization Performance Mainly Stem from High Train-Test Similarity? In *ICLR*, 2024.

Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, and Ludwig Schmidt. Quality Not Quantity: On the Interaction between Dataset Design and Robustness of CLIP. In *NeurIPS*, 2022.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. *arXiv*, 2018.

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 *ICML*, 2021.

Cyrus Rashtchian, Charles Herrmann, Chun-Sung Ferng, Ayan Chakrabarti, Dilip Krishnan, Deqing Sun, Da-Cheng Juan, and Andrew Tomkins. Substance or Style: What Does Your Image Embedding Know? *arXiv*, 2023.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Datasets and Benchmarks Track@NeurIPS*, 2022.

Peiyang Shi, Michael C. Welle, Mårten Björkman, and Danica Kragic. Towards understanding the modality gap in CLIP. In *Workshop@ICLR*, 2023.

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does CLIP know about a red circle? Visual prompt engineering for VLMs. In *ICCV*, 2023.

Junhyuk So, Changdae Oh, Yongtaek Lim, Hoyoon Byun, Minchul Shin, and Kyungwoo Song. Geodesic Multi-Modal Mixup for Robust Fine-Tuning. In *NeurIPS*, 2023.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: Improved Training Techniques for CLIP at Scale. *arXiv*, 2023.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The New Data in Multimedia Research. *Communications of the ACM*, 2016.

Matthew Trager, Pramuditha Perera, Luca Zancato, Alessandro Achille, Parminder Bhatia, and Stefano Soatto. Linear Spaces of Meanings: Compositional Structures in Vision-Language Models. In *ICCV*, 2023.

Vishaal Udandarao. Understanding and Fixing the Modality Gap in Vision-Language Models, 2022. Master's thesis.

Alexander Visheratin. NLLB-CLIP–train performant multilingual image retrieval model on a budget. *arXiv*, 2023.

Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style. In *NeurIPS*, 2021.

Tongzhou Wang and Phillip Isola. Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere. In *ICML*, 2020.

Chenyun Wu and Subhransu Maji. How well does CLIP understand texture? *Workshop@ECCV*, 2022.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying CLIP Data. In *ICLR*, 2024.

Yutaro Yamada, Yingtian Tang, and Ilker Yildirim. When are Lemons Purple? The Concept Association Bias of CLIP. In *EMNLP*, 2023.

Aron Yu and Kristen Grauman. Fine-Grained Visual Comparisons with Local Learning. In *CVPR*, 2014.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. CoCa: Contrastive Captioners are Image-Text Foundation Models. *TMLR*, 2022.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and Why Vision-Language Models Behave like Bags-Of-Words, and What to Do About It? In *ICLR*, 2022.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. LiT: Zero-Shot Transfer with Locked-image text Tuning. In *CVPR*, 2022.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training. In *ICCV*, 2023.

Renrui Zhang, Ziyao Zeng, Ziyu Guo, and Yafeng Li. Can Language Understand Depth? In *International Conference on Multimedia*, 2022.

Chenliang Zhou, Fangcheng Zhong, and Cengiz Öztireli. CLIP-PAE: Projection-Augmentation Embedding to Extract Relevant Features for a Disentangled, Interpretable and Controllable Text-Guided Face Manipulation. In *SIGGRAPH*, 2023.

A Related work

Contrastive multi-modal representation learning. Multi-modal representation learning aims to learn a shared representation space from heterogeneous input modalities. In this work, we focus on VLMs that are trained by contrastive learning. Contrastive learning aims to *align* related data from different modalities and to enforce *uniformity* of the induced distribution on the unit hypersphere (Oord et al., 2018; Wang & Isola, 2020). Following the popular CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021), other works proposed several techniques to improve reproducibility (Cherti et al., 2023) or training efficiency (Zhai et al., 2022; Li et al., 2023d; Zhai et al., 2023; Li et al., 2023b;a; Sun et al., 2023).

Understanding contrastive multi-modal representation learning. Given its success across diverse tasks, recent attention has been directed toward understanding contrastive multi-modal representation learning. Prior work studied the importance of data (Nguyen et al., 2022; Xu et al., 2024), generalization or robustness (Mayilvahanan et al., 2024; Crabbé et al., 2023), analyzed the learned features/representations (Goh et al., 2021; Materzyńska et al., 2022; Rashtchian et al., 2023), or learned (social) biases and capabilities (Agarwal et al., 2021; Yamada et al., 2023; Zhang et al., 2022; Wu & Maji, 2022; Shtedritski et al., 2023; Hamidieh et al., 2023). Other work studied the compositionality of CLIP's embedding space through vector arithmetic (Jia et al., 2021; Couairon et al., 2022; Trager et al., 2023). Specifically, Trager et al. (2023) proposed computing *ideal words* by marginalizing out other factors from captions. While some works found that large models, including VLMs, close the gap to human perception (Geirhos et al., 2021; Li et al., 2023c), another line of works found several failure modes of VLMs (Yuksekgonul et al., 2022; Brody, 2023).

Recent work showed the presence of a *modality gap* and found evidence that it appears due to the cone effect of model initialization and the contrastive learning objective (Liang et al., 2022). Subsequent work explored the influence of the Softmax temperature (Udandarao, 2022; Shi et al., 2023). In our work we fine evidence against the cone effect hypothesis by prior work (Liang et al., 2022) as the sole cause of modality gap and identify the information imbalance between the modalities as a trigger for the modality gap. Further, we approach the question whether the gap is an actual problem by a large-scale analysis of 113 contrastive VLMs.

Orthogonal to the above, other works disentangled the standard InfoNCE (contrastive) loss (Oord et al., 2018) into an alignment and uniformity component (Wang & Isola, 2020). Recent works theoretically analyzed the (multi-modal) contrastive loss and identified the importance of shared task-relevant information, i.e., content, between the modalities (Von Kügelgen et al., 2021; Daunhawer et al., 2023; Liang et al., 2023). We connect the per-sample information imbalance between the modalities to two observed phenomena.

B Experimental setup for Sections 2 to 4

Contrastive vision-language models. Unless otherwise stated, we used CLIP ViT-B/16 (Radford et al., 2021) for our analysis. For our large-scale analyses, we used a total of 113 contrastive VLMs trained across various datasets provided by OpenCLIP (Ilharco et al., 2021; Cherti et al., 2023)[1]. It contains contrastive VLMs, such as OpenAI's CLIP (Radford et al., 2021), CLIP-A (Li et al., 2023b), EVA-CLIP (Sun et al., 2023), CoCa (Yu et al., 2022), NLLB-CLIP (Visheratin, 2023), or SigLIP (Zhai et al., 2023). Note that these models use various backbones, including ResNet (He et al., 2016), ConvNeXt (Liu et al., 2022), or ViT (Dosovitskiy et al., 2020). The models were trained on, e.g., OpenAI's proprietary (400 M) WebImageText dataset (Radford et al., 2021), LAION-400 M, LAION-2 B, LAION-5 B (Schuhmann et al., 2022), Merged-2 B (merge of 1.6 B samples from LAION-2 B and 0.4 B samples from COYO-700 M (Byeon et al., 2022)) (Sun et al., 2023), WebLI (Chen et al., 2023), So-400 M (Alabdulmohsin et al., 2023), MetaCLIP (400 M) (Xu et al., 2024), Conceptual 12 M (Changpinyo et al., 2021), YFCC (15 M) (Thomee et al., 2016), CommonPool-s (max. 12.8 M; refer to Table 3 of Gadre et al. (2023) for the details of filtering), CommonPool-m (max. 128 M), CommonPool-l (max. 1.28 B), CommonPool-xl (max. 12.8 B) (Gadre et al., 2023), or DataPool-s (1.4 M), DataPool-m (14 M), DataPool-l (140 M), DataPool-xl (1 B) (Gadre et al., 2023).

[1] https://github.com/mlfoundations/open_clip

In our analysis, we distinguished between medium- (i.e., dataset size of $\leq$128 M) and large-scale datasets.

Evaluation datasets. We conducted our evaluation of contrastive VLMs on ImageNet (Russakovsky et al., 2015), MS COCO (Lin et al., 2014; Chen et al., 2015), MIT-States (Isola et al., 2015), and UT-Zappos (Yu & Grauman, 2014). The datasets comprise 50000, 25000 (5000 images with 5 captions each), 12995, or 2914 test samples, respectively. ImageNet and MS COCO are standard datasets for evaluation of object recognition or retrieval performance, respectively. We used the standard evaluation protocols to compute accuracy or image retrieval performance. MIT-States consists of 245 objects and 115 adjectives (attributes), while UT-Zappos consists of 12 shoe types with 16 fine-grained states ($\sim$ attributes). For both datasets, we assume that we do not know the object of a respective image and only want to find the adjective or fine-grained state. We considered them as a classification problem, following previous work (Trager et al., 2023). Note that these datasets implicitly assume that the adjectives are mutually exclusive per image. However, this may not be necessarily true, as multiple adjectives or fine-grained states may be present in an image.

For ImageNet, we used the CLIP-style prompts `"a photo of a {obj}"` (Radford et al., 2021) and computed the zero-shot (object) accuracy. For MS COCO, we prepended the prompt `"a photo of"` to the description of each image following Radford et al. (2021) and used R@1 to assess zero-shot image retrieval performance. For MIT-States and UT-Zappos we adopted similar prompts: `"an image of a {attr} object"` and computed the zero-shot attribute accuracy.

C Experimental details for experiments on multi-modal attributes and digits (MAD)

To understand the influence of data and embedding size, we constructed a multi-modal dataset based on the MNIST (LeCun, 1998) variation Morpho-MNIST (Castro et al., 2019) called Multi-modal Attributes and Digits (MAD) with full control over the data-generating process. We adopted the following morphing or warping operations as latent factors (i.e., attributes): altering image thickness, swelling, fractures from (Castro et al., 2019) and added scaling, colors and captions. To generate image captions, we mapped the digit class and latent factors to words and chained them together, e.g., `0-thickening-swelling-fractures-large-blue`. Specifically, we used the following words for digits (`0`, ..., `9`), altering image thickness (`thickening`, `thinning`, `no thickthinning`), swelling (`swelling`, `no swelling`), fractures (`fracture`, `no fracture`), scaling (`large`, `small`), and color (`gray`, `red`, `green`, `blue`, `cyan`, `magenta`, `yellow`). Thus, we have 16 different attributes. Example image-caption pairs are provided in Fig. 4.

To study the impact of missing information in captions, we considered five cases. In each case, the object, i.e., digit, remains consistently present, while we alter the number of attributes mentioned in the captions, ranging from one to five, by randomly selecting them at every batch. This allows us to simulate diverse cases of information imbalance in captions, spanning from cases with a lot of missing information to those with full information. We provide examples below, where we sequentially remove the amount of information within the captions, i.e., fewer latent factors (attributes) are present:

- Full information setting (i.e., digit & all five attributes)
  - `yellow-swelling-thickening-9-large-fracture`
  - `swelling-thickening-6-red-small-fracture`
  - `5-large-yellow-no swelling-fracture-thinning`
- Partial information setting I (i.e., digit & four attributes)
  - `yellow-swelling-thickening-9-large`
  - `swelling-thickening-6-red-small`
  - `5-large-yellow-no swelling-fracture`
- Partial information setting II (i.e., digit & three attributes)
  - `yellow-swelling-thickening-9`

  - `swelling-thickening-6-red`
  - `5-large-yellow-no swelling`
- Partial information setting III (i.e., digit & two attributes)
  - `yellow-swelling-9`
  - `swelling-thickening-6`
  - `5-large-yellow`
- Partial information setting IV (i.e., digit & one attributes)
  - `yellow-9`
  - `swelling-6`
  - `5-large`

Note that while all the latent factors, i.e., digit and all five attributes, are still visible in the image, the caption may only provide partial information, i.e., attributes are missing from the caption.

Model details. We used smaller CLIP models to train on MAD. Specifically, the ViT-based vision backbone comprises 6 layers, each with a dimensionality d of 256 and $\lfloor d/64 \rfloor = 4$ heads. The transformer-based language backbone also comprises 6 layers, each with a dimensionality of 256 and 8 heads. We used a patch size of 7 and set the context length to 8. The vocabulary consists of 28 words, i.e., all the words for digits (10) and attributes (16), as well as a start and end symbol (2).

Training details. We trained all models with a batch size of 128 for 200 epochs with a learning rate warm-up period of 5 epochs. We used AdamW (Loshchilov & Hutter, 2019) as optimizer with cosine annealing learning rate schedule. We always selected the best performing learning rate across 3 learning rates $\{5e-4,\ 5e-5,\ 1e-5\}$ each trained with 3 random seeds. The best learning rate was selected by comparing average accuracies over the ideal word accuracy, average precision and zero-shot accuracy on all attributes and the class label. For all of our results, we report the average over 3 random seeds.

D Further experimental results

Object bias is not caused by word frequency. Fig. 5 shows that the object bias is not caused by the word frequency, i.e., attributes are even more common than objects in the LAION-2B captions.

Object performance improvements transfer to improvements in attribute detection. Fig. 6 shows that improvements on object-related tasks (ImageNet & MS COCO) transfer to improvements in attribute detection (MIT-States & UT-Zappos).

Perfect image-text alignments can close an initial modality gap at model initialization. Fig. 7(left) shows that a model may have a modality gap at model initialization due to the cone effect (Liang et al., 2022) but it can be closed by the contrastive loss under perfect image-text alignments (Fig. 7(right)).

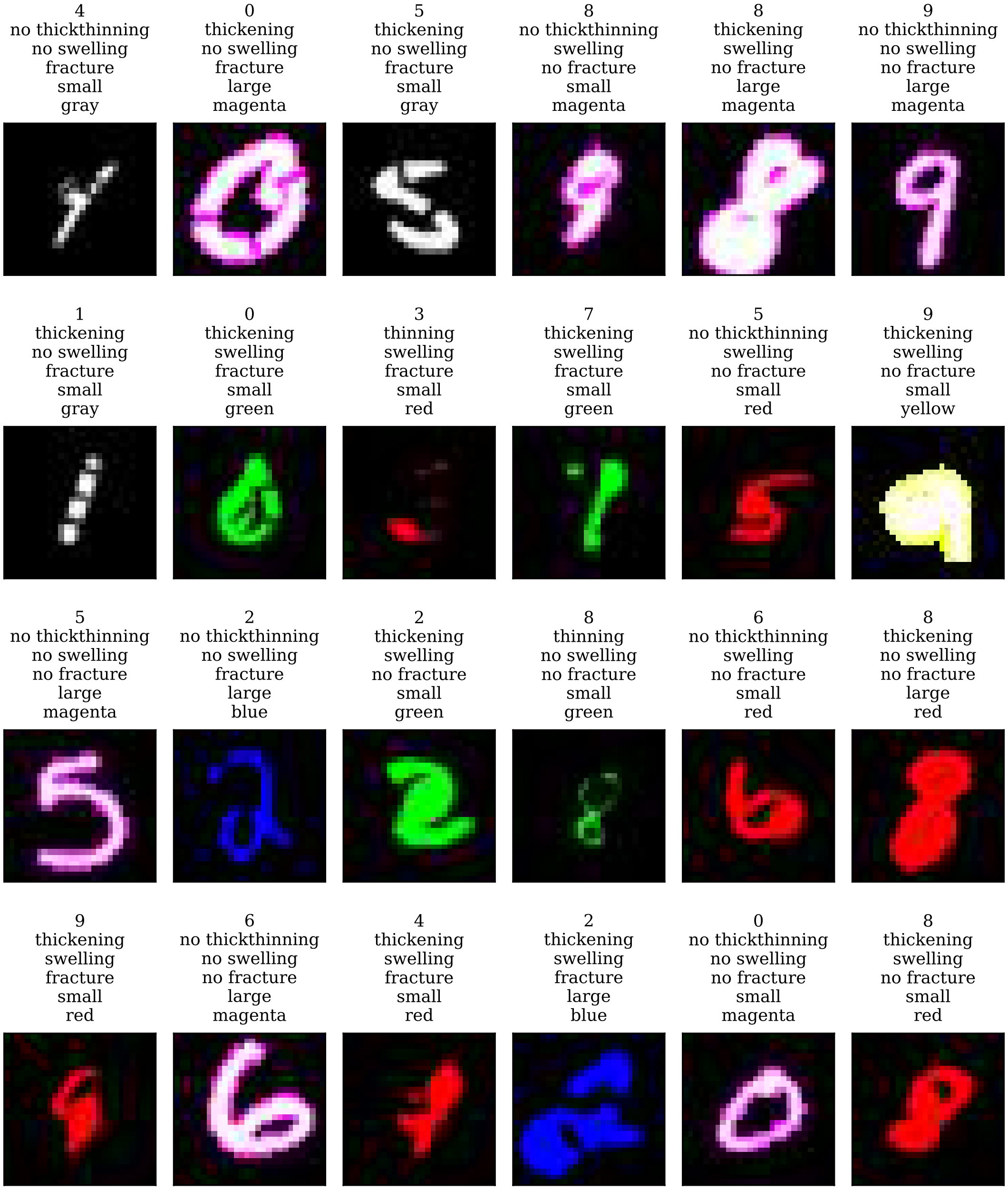


Figure 4: Example image with corresponding caption of our MAD dataset. Note that the words of the captions are shuffled during training. For example, the in the first row and first column shows the digit 4 without altering the thickness, no swelling applied, with fracture augmentation, scaled down and the color gray.

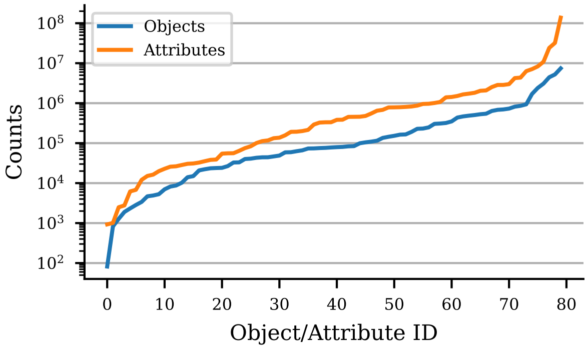


Figure 5: **Object and attribute frequencies** on LAION-2B captions. We used objects and attributes from OVAD (Bravo et al., 2023). Object bias is not caused by the word frequency.

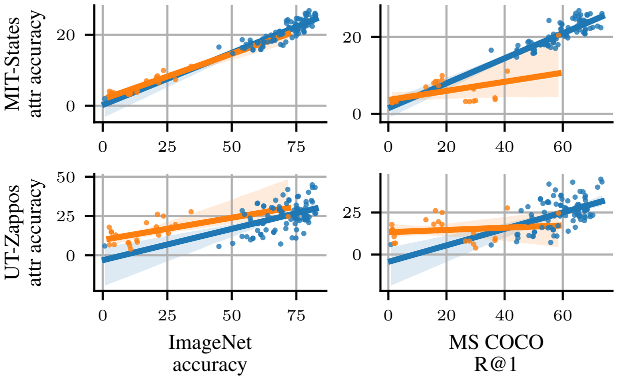


Figure 6: **Relation between object and attribute performance.** There is a strong relation between object and attribute performance, indicating that advances on object performance transfer to attribute performance.

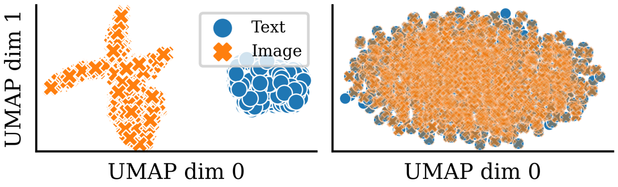


Figure 7: **A contrastive loss with full information can close the modality gap in MAD.** UMAP embeddings after model initialization (left) and after contrastive pre-training (right).