Making LVLMs Look Twice: Contrastive Decoding with Contrast Images

Avshalom Manevich
Bar Ilan University
avshalomman@gmail.com

Reut Tsarfaty
Bar Ilan University
reut.tsarfaty@biu.ac.il

Abstract

Large Vision-Language Models (LVLMs) are becoming increasingly popular for text-vision tasks requiring cross-modal reasoning, but often struggle with fine-grained visual discrimination. This limitation is evident in recent benchmarks like NaturalBench and D3, where closed models such as GPT-4o achieve only 39.6%, and open-source models perform below random chance (25%). We introduce Contrastive decoding with Contrast Images (CoCI), which adjusts LVLM outputs by contrasting them against outputs for similar images (Contrast Images - CIs). CoCI demonstrates strong performance across three distinct supervision regimes: First, when using naturally occurring CIs in benchmarks with curated image pairs, we achieve improvements of up to 98.9% on NaturalBench, 69.5% on D3, and 37.6% on MMVP. Second, for scenarios with modest training data (~5k samples), we show that a lightweight neural classifier can effectively select CIs from similar images at inference time, improving NaturalBench performance by up to 36.8%. Third, for scenarios with no training data, we develop a caption-matching technique that selects CIs by comparing LVLM-generated descriptions of candidate images. Notably, on VQAv2, our method improves VQA performance even in pointwise evaluation settings without explicit contrast images. Our approach demonstrates the potential for enhancing LVLMs at inference time through different CI selection approaches, each suited to different data availability scenarios.

1 Introduction

Large Vision-Language Models (LVLMs) are becoming increasingly popular for text-vision tasks that require reasoning over both modalities. However, they often struggle with fine-grained visual discrimination — that is, the ability to tell two similar yet distinct images apart — a crucial capability for real-world applications such as multimodal search, manufacturing, and robotics. Recent benchmarks have exposed this limitation: on NaturalBench (Li et al., 2024a), which tests visual question answering over closely related images, state-of-the-art closed models like GPT-4o (OpenAI et al., 2024) achieve only 39.6% accuracy. Similarly, on the D3 benchmark (Gaur et al., 2024), which requires describing differences between paired images, open-source models perform below random chance (25%).

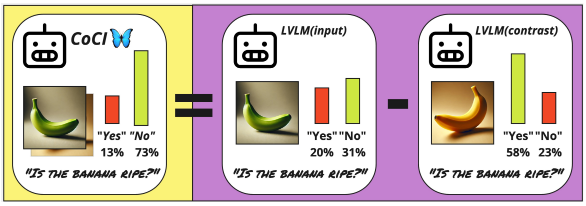

Figure 1: CoCI penalizes target image logits using those from a contrast image, weighted by hyperparameter α.

Efforts to address fine-grained visual discrimination in LVLMs are still under-explored. Current strategies addressing other LVLM shortcomings often rely on fine-tuning with specialized datasets (Wang et al., 2023; Chen et al., 2023; Liu et al., 2024a; Sarkar et al., 2024), multi-step correction pipelines (Yin et al., 2023; Zhou et al., 2023), or inference-time methods (Leng et al., 2023; Manevich and Tsarfaty, 2024; Liu et al., 2024b; Huang et al., 2023). Inference-time methods are particularly appealing as they do not require expensive model training and are less prone to compounding errors that can affect multi-step systems.

Building on the advantages of inference-time methods, we propose Contrastive decoding with Contrast Images (CoCI), an approach specifically designed to improve fine-grained visual discrimination in LVLMs. CoCI penalizes LVLM next-token probabilities with those obtained by feeding a different, contrasting image input (See Figure 1).

We evaluate CoCI across three different supervision regimes. First, using naturally occurring

Proceedings of the 1st Workshop on Multimodal Augmented Generation via Multimodal Retrieval (MAGMaR 2025), pages 65–78

Contrast Images in curated benchmarks like NaturalBench, D3 and MMVP, we demonstrate improvements up to 98.9%, 69.5%, 37.6% respectively. This establishes a performance ceiling for CoCI when ideal CIs are available. For applications where natural CIs are unavailable but training data exists, we show that a lightweight classifier can effectively select CIs from visually similar images at inference time, improving NaturalBench performance by up to 36.5%. In settings without training data, we propose a caption-matching technique that selects CIs at inference time by comparing LVLM-generated descriptions of candidate images.

Experiments with leading LVLMs — Qwen2-VL, LLaVA-OneVision, and Llama 3.2 (Wang et al., 2024a; Li et al., 2024b; Grattafiori et al., 2024) — establish the potential of contrastive decoding strategies with contrastive images for improved multimodal reasoning in real-world tasks.

2 Contrastive Decoding with Contrast Images (CoCI)

We present CoCI, a method to improve LVLM outputs by penalizing token probabilities that are likely under a contrast image. The choice of contrast image is crucial: e.g., when querying about fruit ripeness with an input image of an unripe banana, contrasting against an image of a ripe banana provides strong contrastive signal, while an image of a ripe pear offers weaker contrast and an image of a bus provides no useful signal and may degrade performance. This intuition guides our CI selection strategies across different scenarios. Before formalizing this intuition, we first review key concepts in LVLM text generation.

2.1 Preliminaries: Text Generation in LVLMs

LVLMs extend LLMs by conditioning next-token prediction on both text and images.[1] Generation proceeds by iteratively sampling tokens from the model's predicted distributions until reaching an EOS token or length limit. The LVLM next-token prediction is:

$$\text{LVLM}t(y< t, I) = P(y|y< t, I) \quad \forall y \in \mathcal{V} \quad (1)$$

where $y_{<t}$ is the token prefix, I is the input image, and $\mathcal{V}$ is the model's vocabulary.

2.2 Contrastive Decoding

Following Li et al. (2023), various Contrastive Decoding approaches have emerged (Sennrich et al., 2024; Jin et al., 2024; Phan et al., 2024). We implement CoCI based on Sennrich et al. (2024)'s minimal variant:

$$\text{CoCI}_t(y_{<t}, I, I') = \log\left(P(y|y_{<t}, I) - \alpha P(y|y_{<t}, I')\right) \quad \forall y \in \mathcal{V} \quad (2)$$

CoCI penalizes token probabilities from the target image distribution $P(y|y_{<t}, I)$ with those from the contrast image distribution $P(y|y_{<t}, I')$. The parameter α controls penalty strength.[2]

2.3 Obtaining Contrast Images

We propose three approaches for obtaining CIs:

Naturally occurring CIs. Many tasks naturally provide pairs of images that can serve as contrast images (CIs). For instance, a home assistant robot searching for "the blue ceramic mug with a chip on the handle" needs to distinguish between similar cups to find the exact match. We evaluate this scenario using LVLM benchmarks with curated image pairs designed to test fine-grained discrimination capabilities. These paired images serve as natural CIs in our experiments.

Classifier-obtained CIs. For cases without natural CIs, we train an MLP classifier to select them during inference. Given LVLM L and training triplets $\langle q, I, I' \rangle$ (binary question and image pairs with different answers), we: (a) Extract LVLM hidden states $h_{q,i} \in R^{d_L}$ per image-question pair. (b) Concatenate states for image pairs: $h_{q,i,i'} \in R^{2*d_L}$. (c) Create negative samples using the j least similar images from top-k similar images to I in dataset D[3]. (d) Train a three-layer MLP classifier.[4] We train on NaturalBench (60% split) augmented with GPT-4-generated question paraphrases. At inference, we select the CI maximizing classifier score among k most similar images.[5]

Caption-matched CIs. For scenarios without training data, we select CIs by comparing LVLM-generated image descriptions. Given an input image, we (a) Retrieve k similar images[6]. (b) Generate LVLM descriptions for all $k + 1$ images. (c)

[1]In this work, we focus on single image inputs.

[2]We use $\alpha = 0.5$ for VQA and $\alpha = 0.8$ for open-ended generation.

[3]$j = 5$, $k = 100$. Using flickr30k (Young et al., 2014) and open-clip (Ilharco et al., 2021; Cherti et al., 2023; Radford et al., 2021a; Schuhmann et al., 2022) with cosine similarity.

[4]See appendix A.1 and A.3 for implementation details.

[5]See table 2 for k value comparisons. Inference uses identical retrieval setup as training.

[6]We set $k = 5$ without tuning.

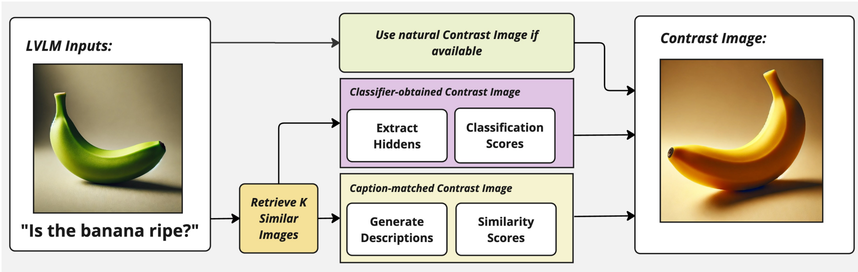

Figure 2: Illustration of the approaches we explore for obtaining a Contrast Image (CI).

Embed descriptions using a text encoder. (d) Select the image whose description is most similar to the input image's description

2.4 Research Hypothesis

We test whether: (a) Contrastive decoding with CIs improves LVLM fine-grained reasoning, (b) A lightweight classifier trained on LVLM hidden states can effectively select CIs, and (c) Images with similar LVLM descriptions can serve as CIs.

3 Experiments

We evaluate CoCI using three leading LVLMs[7] on four benchmarks, three specifically targeting fine-grained visual discrimination:

NaturalBench (Li et al., 2024a) evaluates similar image discrimination through yes/no and multiple-choice questions, with different answers for paired images. The benchmark contains 1900 image pairs (two questions per pair), split into train (60%), dev (20%), and test (20%) sets. We measure image accuracy (both questions correct), question accuracy (per-question), and group accuracy (all four image-question combinations correct).

MMVP (Multimodal Visual Patterns) (Tong et al., 2024) evaluates visual difference detection through multiple-choice questions on 150 image pairs. Each pair differs in specific visual aspects (object state, position, or orientation). Success requires correct answers for both images in a pair.

D3 (Detect, Describe, Discriminate) (Gaur et al., 2024) assesses models' ability to generate discriminative descriptions between similar images across 247 pairs. We adapt D3 for CoCI by treating it as a single-input task, generating separate descriptions per image. Evaluation follows the original self-retrieval protocol, measuring whether an image-text encoder correctly matches descriptions to their images.

[7]See appendix A.2 for details on the checkpoints we used.

Model	Method	D3 (self-ret.)	MMVP (acc.)	NB (g-acc.)	VQAv2 (acc.)
Qwen2-VL	Baseline	30.8	46.0	30.8	72.66
	CoCI_{CAP}	34.8	48.7	31.3	**74.33**
	CoCI_{NAT}	**52.2**	**63.3**	**46.6**	-
LLaVA-OV	Baseline	25.1	52.7	28.2	61.66
	CoCI_{CAP}	31.6	57.3	31.6	**73.66**
	CoCI_{NAT}	**38.1**	**66.7**	**56.1**	-
Llama 3.2	Baseline	28.7	39.3	21.1	58
	CoCI_{CAP}	33.6	41.3	22.4	58
	CoCI_{NAT}	**35.6**	**43.3**	**29.2**	-

Table 1: CoCI performance comparison with provided CIs across benchmarks, with natural CIs (CoCI_{NAT}) and caption-matched CIs (CoCI_{CAP}).

VQAv2 (Goyal et al., 2017) serves as our general-purpose visual question answering benchmark. While not focused on fine-grained discrimination, we include it to demonstrate CoCI's broader applicability. We evaluate on 300 validation set image-question pairs using exact match accuracy.

4 Results and Discussion

In Table 1 we can see that using natural CIs yields substantial improvements: up to 21.4 points on D3 (Qwen), 17.3 points on MMVP (LLaVA), and 27.9 points on NaturalBench (LLaVA). Caption-matched CIs show moderate but consistent gains, particularly on D3 where LLaVA improves from 25.1% to 31.6%, suggesting that contrasting against images with similar captions effectively guides visual discrimination. CoCI with caption matching improves performance on VQAv2 for two of the three tested models while maintaining baseline performance for Llama 3.2, demonstrating that CoCI enhances general-purpose VQA abilities beyond fine-grained visual discrimination tasks.

Throughout our experiments, Llama exhibits different behavior compared to other models - showing lower performance and reduced responsiveness

Model	Method	Q-acc	I-acc	Acc	G-acc
Qwen2-VL	Baseline	55.3	59.3	76.8	30.8
	$Cls_{k=4}$	55.5	58.8	76.4	32.1
	$Cls_{k=8}$	56.3	58.9	76.7	32.4
	$Cls_{k=16}$	57.4	60.1	77.2	33.7
	$Cls_{k=32}$	57.8	60.1	77.4	**34.2**
	$Cls_{k=64}$	**58.2**	**60.8**	**77.9**	33.9
LLaVA-OV	Baseline	53.8	56.1	74.6	28.2
	$Cls_{k=4}$	59.2	59.6	77.6	35.3
	$Cls_{k=8}$	57.8	60.1	77.5	34.5
	$Cls_{k=16}$	57.6	58.7	77.0	33.4
	$Cls_{k=32}$	**60.3**	**62.1**	**78.5**	**38.4**
	$Cls_{k=64}$	59.7	62.1	78.2	37.6
Llama 3.2	Baseline	46.3	50.5	71.8	21.1
	$Cls_{k=4}$	**49.2**	**52.8**	**73.2**	**23.2**
	$Cls_{k=8}$	49.1	52.2	73.1	21.8
	$Cls_{k=16}$	48.8	52.4	73.1	22.4
	$Cls_{k=32}$	49.9	52.5	73.7	22.1
	$Cls_{k=64}$	49.7	52.5	73.6	22.1

Table 2: CoCI accuracy metrics on the NaturalBench test set with CIs chosen using a lightweight classifier. $k = j$ denotes the classifier ran on the j most similar images to the input image.

to our methods. This pattern is evident in Table 2, where Qwen and LLaVA's performance improves with larger candidate pools (k), peaking around k=32, while Llama performs best with small pools (k=4). This behavior could be attributed to two factors: First, while the hyperparameters worked well for Qwen and LLaVA, they may not be optimal for Llama without model-specific tuning. Second, Llama's architectural differences, particularly its use of cross-attention, could lead to different behaviors in our contrastive decoding context. While exploring these architecture-specific considerations could be valuable, it is beyond the scope of this work.

In NaturalBench, G-Acc shows particularly strong improvement with natural CIs as it requires consistency across all image-question combinations. This pattern persists with classifier-selected CIs, where G-Acc improves by up to 10.2 points while other metrics show modest gains. The substantial gap between natural CIs and other methods suggests that classifier-selected and caption-matched CIs, while beneficial, don't yet capture all aspects that make natural pairs effective. [8]

5 Related Work

Inference-time methods for enhancing multimodal reasoning. Recent work has focused on hallucination reduction through confidence-based adjustments (Huo et al., 2024), semantic references (Yang et al., 2024), and contrastive decoding with perturbed inputs (Leng et al., 2023; Manevich and Tsarfaty, 2024). Our work extends these approaches to fine-grained visual discrimination.

Alignment and grounding in LVLMs. Prior work has enhanced visual-textual alignment through object-level synthesis (Wang et al., 2024b), targeted fine-tuning (Lu et al., 2024), and dataset construction (Li et al., 2024c). While these methods improve foundational capabilities, they don't directly address fine-grained discrimination.

Contrastive examples in multimodal models. CLIP (Radford et al., 2021b) established contrastive learning for modality alignment. Recent works leverage contrast pairs: (Le et al., 2023) and (Zhang et al., 2024) generate synthetic datasets using text-to-image models, while (Abbasnejad et al., 2020) and (Zhou et al., 2024) use contrastive examples to address dataset biases. Unlike these approaches requiring data generation or training, our method operates at inference time.[9]

6 Conclusion

We introduced Contrastive decoding with Contrast Images (CoCI), demonstrating its effectiveness in improving LVLMs' fine-grained visual discrimination capabilities in both VQA and long-form generation tasks. While naturally occurring contrast pairs yielded the strongest gains, both classifier-based and caption-matching approaches provide meaningful improvements without requiring dataset curation or model training. We validated the generality of our method through experiments with caption-based contrast selection, showing that CoCI does not rely on pre-curated pairs but can leverage them when available. Notably, CoCI improves performance even on tasks that don't explicitly measure fine-grained discrimination.

Our results show that contrastive decoding algorithms, when combined with strategic contrast image selection, improve LVLMs' ability to make fine-grained distinctions and their overall VQA abilities, opening new avenues for improving multimodal reasoning through inference-time techniques.

[8] See appendix A.3 for ablation tests with different CI selection strategies.

[9] Classifier-selected CIs require minimal preprocessing compared to model finetuning or dataset curation.

7 Limitations

CoCI has several limitations worth noting. While we demonstrate its effectiveness with classifier-based and caption-matching approaches, the substantial performance gap between natural and automatically selected CIs indicates significant headroom for finding more effective contrast images. We tested simple selection methods to establish the viability of the approach, leaving the exploration of more sophisticated CI selection strategies to future work. Additionally, our evaluation focuses primarily on VQA and self-retrieval protocols; exploring additional evaluation methods could reveal other aspects of how CoCI affects LVLM generations.

The method introduces additional computation at inference time, running the LVLM twice per generation step and requiring CI selection overhead. While this aligns with the growing trend of leveraging test-time compute for improved performance, the current implementation could be optimized. Future work could explore more efficient implementations of contrastive decoding and investigate fusing operations like hidden state extraction with the generation procedure to reduce computational overhead.

Our implementation uses Flickr30k as the image database for CI selection - using larger, more diverse image collections could improve performance. Alternative image retrieval models and similarity scoring methods could also enhance CI selection. Additionally, our approach does not address cases where multiple contrasts might be informative - we only use a single contrast image, while some scenarios might benefit from multiple contrasting viewpoints.

The experiments use a fixed contrastive weight (α) across tasks within each category (VQA/generation). A more nuanced approach to setting this parameter, dynamically per sample or per token, based on the specific input or task, could yield better results.

While CoCI improves visual discrimination, it could potentially amplify biases present in contrast image databases or introduce new failure modes when inappropriate contrast images are selected. These risks should be carefully evaluated before deployment in sensitive applications.

Finally, our experiments focus exclusively on English-language benchmarks. Extending CoCI to multilingual settings and investigating how contrastive decoding approaches perform across different languages represents an important direction for future research.

References

Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. 2020. Counterfactual vision and language learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10041–10051.

Zhiyang Chen, Yousong Zhu, Yufei Zhan, Zhaowen Li, Chaoyang Zhao, Jinqiao Wang, and Ming Tang. 2023. Mitigating hallucination in visual language models with visual supervision.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829.

Manu Gaur, Darshan Singh S, and Makarand Tapaswi. 2024. Detect, describe, discriminate: Moving beyond vqa for mllm evaluation.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park,

Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal

Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Qidong Huang, Xiao wen Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Neng H. Yu. 2023. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13418–13427.

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. 2024. Self-introspective decoding: Alleviating hallucinations for large vision-language models.

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. 2021. Openclip. If you use this software, please cite it as below.

Jing Jin, Houfeng Wang, Hao Zhang, Xiaoguang Li, and Zhijiang Guo. 2024. DVD: Dynamic contrastive decoding for knowledge amplification in multi-document question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4624–4637, Miami, Florida, USA. Association for Computational Linguistics.

Tiep Le, Vasudev Lal, and Phillip Howard. 2023. Coco-counterfactuals: Automatically constructed counterfactual examples for image-text pairs.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2023. Mitigating object hallucinations in large vision-language models through visual contrastive decoding.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. 2024a. Naturalbench: Evaluating vision-language models on natural adversarial samples.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024b. Llava-onevision: Easy visual task transfer.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization.

Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Van Tu Vu, Zhida Huang, and Tao Wang. 2024c. Groundinggpt:language enhanced multi-modal grounding model.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024a. Mitigating hallucination in large multi-modal models via robust instruction tuning.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. 2024b. Reducing hallucinations in vision-language models via latent space steering.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Junyu Lu, Dixiang Zhang, Songxin Zhang, Zejian Xie, Zhuoyang Song, Cong Lin, Jiaxing Zhang, Bingyi Jing, and Pingjian Zhang. 2024. Lyrics: Boosting fine-grained language-vision alignment and comprehension via semantic-aware visual objects.

Avshalom Manevich and Reut Tsarfaty. 2024. Mitigating hallucinations in large vision-language models (LVLMs) via language-contrastive decoding (LCD). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6008–6022, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin

Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. Gpt-4o system card.

Phuc Phan, Hieu Tran, and Long Phan. 2024. Distillation contrastive decoding: Improving llms reasoning with contrastive decoding and distillation.

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. 2021a. Learning transferable visual models from natural language supervision. In *ICML*.

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. 2021b. Learning transferable visual models from natural language supervision.

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö. Arık, and Tomas Pfister. 2024. Data-augmented phrase-level alignment for mitigating object hallucination.

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. 2022. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-*

sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track.

Rico Sennrich, Jannis Vamvas, and Alireza Mohammadshahi. 2024. Mitigating hallucinations and off-target machine translation with source-contrastive and language-contrastive decoding.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms.

Lei Wang, Jiabang He, Shenshen Li, Ning Liu, and Ee-Peng Lim. 2023. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites.

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. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution.

Wei Wang, Zhaowei Li, Qi Xu, Linfeng Li, YiQing Cai, Botian Jiang, Hang Song, Xingcan Hu, Pengyu Wang, and Li Xiao. 2024b. Advancing fine-grained visual understanding with multi-scale alignment in multi-modal models.

Dingchen Yang, Bowen Cao, Guang Chen, and Changjun Jiang. 2024. Pensieve: Retrospect-then-compare mitigates visual hallucination.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2023. Woodpecker: Hallucination correction for multimodal large language models.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Jianrui Zhang, Mu Cai, Tengyang Xie, and Yong Jae Lee. 2024. Countercurate: Enhancing physical and semantic visio-linguistic compositional reasoning via counterfactual examples.

Baohang Zhou, Ying Zhang, Kehui Song, Hongru Wang, Yu Zhao, Xuhui Sui, and Xiaojie Yuan. 2024. MCIL: Multimodal counterfactual instance learning for low-resource entity-based multimodal information extraction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11101–11110, Torino, Italia. ELRA and ICCL.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models. *ArXiv*, abs/2310.00754.

A Appendix

A.1 Lightweight Classifier Implementation Details

Below is the PyTorch code of the lightweight classifier.

```
class Classifier(torch.nn.Module):
    def __init__(self, input_dim: int):
        super(Classifier, self).__init__()
        # factor of 2 due to concatentaion of target and candidate features
        self.linear1 = torch.nn.Linear(input_dim * 2, input_dim)
        self.linear2 = torch.nn.Linear(input_dim, input_dim)
        self.linear3 = torch.nn.Linear(input_dim, 1)
        self.dropout = torch.nn.Dropout(p=0.3)

    def forward(self, x) -> torch.Tensor:
        x = self.dropout(self.linear1(x))
        x = F.relu(x)
        x = self.dropout(self.linear2(x))
        x = F.relu(x)
        x = self.linear3(x)
        return x
```

We trained a classifier per tested LVLM, all with the following parameters, using the AdamW (Loshchilov and Hutter, 2019) optimizer.

```
batch_size=256
num_epochs=13
learning_rate=3e-4
weight_decay=1e-6
```

A.2 LVLM Checkpoints Tested

The following are the LVLM checkpoints we tested CoCI with:

```
Qwen/Qwen2-VL-7B-Instruct
llava-hf/llava-onevision-qwen2-7b-ov-hf
meta-llama/Llama-3.2-11B-Vision-Instruct
```

We used *laion/CLIP-ViT-L-14-DataComp.XL-s13B-b90K* as the open-clip model for both image and text encoding throughout this work.

A.3 Effect of Choosing a Contrast Image on NaturalBench Performance

Method	Setting	Q-acc	I-acc	Acc	G-acc
CoCI ablations	Baseline	51.6	55.4	75.1	25.6
	CI ←Random (out of top-5 most similar to input)	49.6	52.1	73.8	23.2
	CI ←Natural	**71.8**	**70.8**	**84.3**	**51.6**
	CI ←Most similar to input	49.7	52.5	73.6	23.9
	CI ←Most similar to Natural	60.3	60.7	78.9	35.0
	CI ←Least similar to Natural	46.7	48.9	72.6	21.8
Classifier	$k = 4$	51.7	54.3	74.5	26.6
	$k = 8$	53.0	55.4	75.3	26.6
	$k = 16$	54.3	56.8	76.1	29.2
	$k = 32$	52.2	54.6	75.1	25.8
	$k = 64$	51.8	53.9	74.7	26.3
	$k = 100$	52.1	54.1	74.8	25.5
Classifier$_{\text{+augmentations}}$	$k = 4$	52.0	54.3	74.6	27.1
	$k = 8$	52.8	55.9	75.0	27.9
	$k = 16$	54.5	57.8	76.1	29.2
	$k = 32$	54.9	58.2	75.9	30.0
	$k = 64$	54.7	57.9	76.1	30.3
	$k = 100$	54.7	58.0	76.1	30.0

Table 3: CoCI performance on the NaturalBench dev set with different CI selection methods, using Qwen2-VL. Classifier$_{\text{+augmentations}}$ indicates training data augmentation with GPT-4o paraphrased questions and standard image augmentations. Using natural CIs provides the strongest performance gains, with a 26-point improvement in group accuracy over baseline (51.6% vs 25.6%). Selecting CIs by similarity to natural CIs improves performance significantly (35.0% G-acc), while using the least similar images performs worse than baseline (21.8%), validating the importance of CI selection strategy. Random CI selection hurts performance (23.2% G-acc) even when restricted to similar images, highlighting that similarity alone is insufficient. Training with augmented data provides modest but consistent improvements across all metrics, with G-acc increasing by about 4 points compared to the non-augmented classifier. The augmented classifier also demonstrates more robust performance, maintaining consistent scores across different k values compared to the higher variance seen in the non-augmented version.

A.4 Examples of Image Pairs and Their Descriptions from D3

The following pairs of images with their respective captions are sampled from D3 results, obtained using CoCI with Natural Contrast Images and Qwen 7B.

Baseline: The image shows a person holding a banana in front of their mouth. The person is wearing a green shirt and is standing against a plain wall. The banana is positioned in such a way that it appears to be covering the person's mouth, creating a humorous or playful effect. The background is simple and uncluttered, which helps to focus attention on the person and the banana.
CoCI: The image shows a person holding a banana close to their face. The person appears to be making an exaggerated facial expression, with their eyes wide open and eyebrows raised. The background is a plain, light-colored wall.

Baseline: The image shows a person holding a banana in front of their mouth, creating the illusion of a smile. The person is wearing a green shirt and is standing against a plain wall. The banana is positioned in such a way that it covers the lower part of the person's face, mimicking a smile. The lighting in the image is even, and the background is simple and uncluttered.
CoCI: The image depicts a man with a light complexion wearing a green shirt. He is leaning against a wall, which appears to be a combination of light and dark shades, possibly indicating a corner or a wall with different sections. The man is holding a banana in his right hand, and he has placed the banana horizontally across his mouth, creating a playful expression that resembles a mustache. The banana has some dark spots on it, suggesting it is ripe. The background is relatively plain, with no visible decorations or objects, focusing the viewer's attention on the man and the banana.

Figure 3: Comparison of baseline and CoCI captions. The baseline captions are very similar and provide no information that allows to discriminate between the images, while the left CoCI caption mentions the raised eyebrows.

Baseline: The image depicts the iconic Elizabeth Tower, commonly known as Big Ben, located in London, England. The tower is a prominent example of Gothic Revival architecture and is part of the Palace of Westminster. The photograph captures the upper portion of the tower, focusing on its clock face and the ornate details of its structure.
CoCI: The image showcases a close-up view of a prominent clock tower, which is characterized by its Gothic architectural style. The tower is adorned with intricate details and features two large, circular clock faces, one on each visible side of the structure. Both clocks are encased in ornate frames, which are embellished with decorative patterns and motifs.

Baseline: The image depicts the iconic Elizabeth Tower, commonly known as Big Ben, located in London, England. The tower is a prominent example of Gothic Revival architecture and is part of the Palace of Westminster. The clock face is visible in the center of the tower, displaying the time. The clock is encased in a decorative frame with intricate designs, and the hands of the clock are clearly visible. The tower is topped with a spire and a weathervane, which adds to its historical and architectural significance. The sky in the background is clear, providing a sharp contrast to the detailed stonework of the tower.
CoCI: The image depicts the iconic clock tower known as Big Ben, located in London, England. The tower is part of the Palace of Westminster and is one of the most recognizable landmarks in the world.

Figure 4: Comparison of baseline and CoCI captions. The left CoCI caption mentions the two visible clock faces.

Baseline: The image depicts a person engaged in windsurfing, a water sport that combines elements of surfing and sailing. The individual is standing on a windsurfing board, which is partially submerged in the water. ... The individual appears to be in a standing position, holding onto the sail with both hands, indicating that they are actively maneuvering the board.
CoCI: The image depicts a person engaging in the water sport of windsurfing near a beach. The individual is wearing a black wetsuit and appears to be holding onto the sail of the windsurfing board ... The person appears to be standing in shallow water, with their upper body and arms visible above the waterline.

Baseline: The image depicts a person engaged in windsurfing, an activity that combines elements of surfing and sailing. The windsurfer is standing on a surfboard, which is partially submerged in the water ... The individual's posture suggests they are actively maneuvering the sail, likely using their body weight and the wind to control the direction and speed of the board.
CoCI: The image depicts a windsurfer riding a wave in the ocean. The windsurfer is captured from the side, showing their body leaning back as they navigate the wave ... The windsurfer is wearing dark clothing and appears to be in motion, with the water splashing around their feet as they ride the wave. ...

Figure 5: Comparison of baseline and CoCI captions. The CoCI captions describe the state of the windsurfer: standing in the shallow water vs. riding the wave.