Is CLIP Cross-Eyed? Revealing and Mitigating Center Bias in the CLIP Family

Oscar Chew[1], **Hsiao-Ying Huang**[2], **Kunal Jain**[1], **Tai-I Chen**[2,3], **Khoa D. Doan**[4], **Kuan-Hao Huang**[1]
[1]Texas A&M University, [2]National Taiwan University, [3]ASUS, [4]VinUniversity
oscarchew@tamu.edu

Abstract

Recent research has shown that contrastive vision-language models such as CLIP often lack a fine-grained understanding of visual content. While a growing body of work has sought to address this limitation, we identify a distinct failure mode in the CLIP family, which we term *center bias*, that persists even in recent model variants. Specifically, CLIP tends to disproportionately focus on the central region of an image, overlooking important objects located near the boundaries. This limitation is fundamental as failure to recognize relevant objects makes it difficult to perform any sophisticated tasks that depend on those objects. To understand the underlying causes of the limitation, we conduct analyses from both representation and attention perspectives. Using interpretability methods, i.e., embedding decomposition and attention map analysis, we find that relevant concepts especially those associated with off-center objects vanish from the model's embedding in the final representation due to information loss during the aggregation of visual embeddings, particularly the reliance on pooling mechanisms. Finally, we show that this bias can be alleviated with training-free strategies such as visual prompting and attention redistribution by redirecting models' attention to off-center regions. Our code is available at https://github.com/lab-flair/vlm-center-bias.

1 Introduction

Contrastive vision–language models (VLMs) such as CLIP (Radford et al., 2021) have become a foundational component in multimodal retrieval and generative systems. Despite their wide applications, recent studies have shown that CLIP and its variants often lack a fine-grained understanding of visual content. They often rely on a coarse understanding or spurious cues, failing to capture detailed object attributes, and exhibit "bag-of-words" behavior, struggling to accurately bind attributes to their corresponding objects (Yuksekgonul et al., 2023; Hsieh et al., 2023; Dumpala et al., 2024; Tong et al., 2024).

Beyond the inability to correctly associate present attributes, we observe a more critical failure mode: concepts can completely vanish from the model's embedding depending on where they appear in the image. Drawing inspiration from human vision research (Tseng et al., 2009; Bindemann, 2010; Borji et al., 2011), which identifies a strong *center bias* in gaze behavior driven by photographer bias and human viewing strategies, we examine whether analogous biases emerge in the CLIP family. As illustrated in Figure 1, when an object is placed away from the center, the model may fail to recognize it altogether even when it is clearly visible. In this example, although both a chair (center) and a pot (off-center) are present, the model assigns high confidence to "a chair" while failing to capture the off-center object. Motivated by this observation, we reveal that CLIP tends to disproportionately focus on the central region of an image, systematically overlooking important objects located near the boundaries.

To systematically study this phenomenon, we evaluate model performance on both a repurposed real-world spatial relation dataset and a family of controlled synthetic datasets,

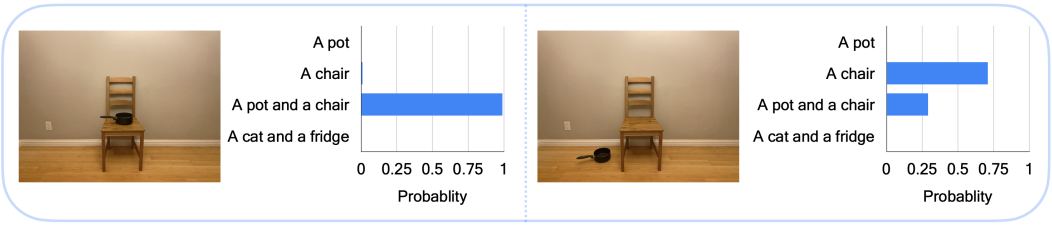

Figure 1: A representative example from **What'sUp**. When both objects are centrally aligned (left), CLIP correctly assigns high probability to "a pot and a chair". However, when the pot is moved off-center (right), CLIP disproportionately focuses on the central chair, leading to a sharp drop in probability for the correct answer and a preference for "a chair."

demonstrating a consistent performance degradation on off-center objects across diverse model variants, including the latest advancements (Monsefi et al., 2025; Zhao et al., 2025). Furthermore, we analyze the underlying causes of this bias from both representation and attention perspectives. Through embedding decomposition and attention map analysis, we find that relevant concepts associated with off-center objects are present in representations of other visual tokens but are not effectively captured in the final embedding. This indicates that the bias stems primarily from information loss during the aggregation into the `[CLS]` token, rather than an inherent lack of available visual information.

Finally, having identified the mechanism behind this information loss, we demonstrate that this bias can be alleviated using training-free strategies. By employing visual prompting and attention redistribution, we can redirect the models' attention to off-center regions and recover overlooked concepts without retraining or modifying model parameters.

We summarize our contributions as follows:

- We identify and quantify center bias in the CLIP family, showing consistent performance degradation on off-center objects across diverse datasets and model variants.
- We provide analyses from both representation and attention perspectives, revealing the underlying mechanisms by which CLIP exhibits center bias.
- We demonstrate that simple strategies, such as visual prompting and attention redistribution, can effectively alleviate center bias.

2 Related Works

2.1 Contrastive Vision-Language Models

Contrastive VLMs have emerged as a fundamental paradigm for aligning visual and textual representations. CLIP (Radford et al., 2021) learns a shared embedding space by contrasting matched image-text pairs against mismatched ones, enabling strong zero-shot transfer across a wide range of tasks. Subsequent works have improved this framework from different perspectives such as scaling to larger datasets and models (Cherti et al., 2023; Chen et al., 2024), integrating contrastive learning with additional pretraining tasks (Yu et al., 2022; Fang et al., 2024), and adopting alternative training objectives (Zhai et al., 2023).

2.2 Robust and Fine-grained Perception of VLMs

Recent work has identified limitations in the fine-grained perception capabilities of the CLIP family. They show that CLIP often relies on coarse or spurious cues, failing to capture detailed object attributes and subtle visual distinctions (Yuksekgonul et al., 2023; Rahmanzadehgervi et al., 2024; Adila et al., 2024; Wang et al., 2024; Zhang et al., 2024; Tong et al., 2024; Koishigarina et al., 2026). To address this, several methods have been introduced to improve fine-grained understanding. NegCLIP (Yuksekgonul et al., 2023) incorporates

hard negative examples to encourage better discrimination, while DetailCLIP (Monsefi et al., 2025) and SuperCLIP (Zhao et al., 2025) augment contrastive learning with dense predictions or other auxiliary objectives to enhance sensitivity to fine-grained features.

In this work, we show that even recent CLIP variants continue to exhibit a systematic *center bias*, indicating that improvements in fine-grained perception do not necessarily translate to better spatial coverage (Section 3.1).

2.3 Center Bias in Human Vision

Research in human vision has identified a strong center bias in gaze behavior, where observers tend to fixate near the center of visual scenes (Tseng et al., 2009; Bindemann, 2010; Borji et al., 2011). This bias arises primarily from two factors: *photographer bias*, where objects of interest are preferentially placed near the center of images, and *viewing strategy*, where observers adopt a heuristic of initially attending to the center.

In this work, we draw inspiration from these findings and investigate whether similar biases emerge in contrastive vision-language models. In particular, we examine whether CLIP models exhibit a form of center bias analogous to human viewing strategy and dataset-driven biases, and whether this bias affects their ability to recognize objects outside the central region.

3 Revealing Center Bias in CLIP

In this section, we investigate the presence and extent of center bias in CLIP-based models. First, we repurpose the **What'sUp** dataset (Kamath et al., 2023) and introduce a family of synthetic datasets, **GRID**, to isolate the effect of object position on model performance. We then establish a metric for center bias and benchmark a wide-range of CLIP variants. We hypothesize that despite advances in model scale, architecture, and training objectives, contrastive models will exhibit a significant and systematic performance degradation when recognizing objects placed outside the center of an image.

3.1 Experiment Setup

Datasets First, we use **What'sUp** (Kamath et al., 2023) subset A, where a single object is positioned relative to another object (e.g., on, left of, right of, or under a table, chair, or armchair). This dataset provides a real-world testbed for vision-language models. To connect this dataset with our study of positional bias, we partition the examples into two groups: **center** and **off-center**. We treat *"on"* relations as the **center** set, since the primary object typically appears near the center of the image, while *"left of," "right of,"* and *"under"* relations form the **off-center** set, where the primary object is more likely to be displaced from the center. We provide further justification for the validity of this setup in Appendix B. Unlike the original What'sUp task, which requires predicting spatial relations, we instead consider a simpler recognition-based question: *"What is in this image?"*. Surprisingly, we find that CLIP cannot reliably answer this easier question. The answer candidates include: (1) the primary object, (2) the supporting object (e.g., chair, table, or armchair), (3) both objects (the ground-truth answer), and (4) a distractor with a token length matched to the ground-truth answer. If CLIP struggles to answer such a simple recognition question, its predictions on more challenging spatial reasoning tasks may be unreliable.

Both object position and size affect whether a model can perceive objects in an image. To isolate the effect of position, we construct a family of synthetic datasets, **GRID**, while controlling for object size. These datasets are built from standard image classification benchmarks, including CIFAR-10 (Krizhevsky & Hinton, 2009), Fashion-MNIST (Xiao et al., 2017), and Food-101 (Bossard et al., 2014). Each GRID sample consists of a larger canvas divided into a $k \times k$ grid, where a single object image, sized s=1/3/5 times the patch size, is placed in the grid. The backgrounds of the grids are chosen from a texture database (Cimpoi et al., 2014). We define two subsets: (1) **GRID-center**, where the object is always placed in the central cell, and (2) **GRID-off-center**, where the object is randomly placed in the outer

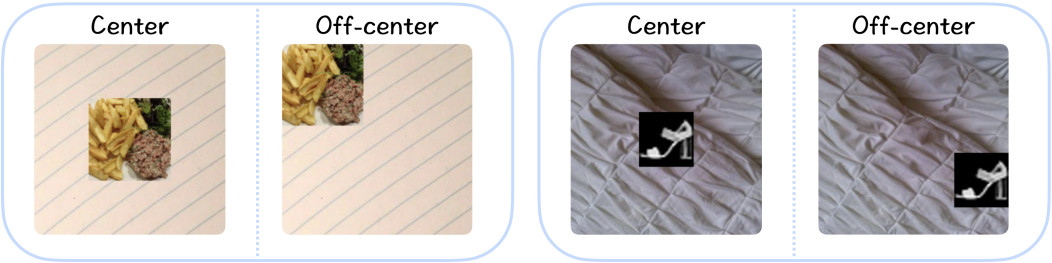

Figure 2: Examples of center and off-center configurations in the **GRID** dataset.

ring of the grid. This design allows us to isolate the effect of object position while keeping all other factors (e.g., object identity, scale, and background) consistent. The examples and construction details of **GRID** are provided in Figure 2 and Appendix A.

Image preprocessing We follow the standard preprocessing steps used in OpenCLIP. Specifically, every image undergoes the same procedure: resizing the shortest side to the target resolution (e.g., 224 pixels), center-cropping the image to (224×224), converting it to a tensor, and normalizing the tensor. This preprocessing does not affect GRID because all **GRID** images are already (224×224) squares. We also manually inspected every image in **What'sUp** after preprocessing and confirmed that the objects in all images remain visible.

Metrics We evaluate model performance using classification accuracy. In particular, we measure accuracy separately on the **center** and **off-center** subsets. *center bias* is quantified as the performance gap between them. Our primary goal is to assess whether model performance is sensitive to object position. An ideal model without center bias should achieve high accuracy on both subsets while exhibiting a small performance gap between center and off-center cases.

Models We study a diverse set of CLIP-based models spanning different architectures, training objectives, scales, and input resolutions. Our evaluation includes OpenAI CLIP (Radford et al., 2021), OpenCLIP (Cherti et al., 2023), CoCa (Yu et al., 2022), SigLIP (Zhai et al., 2023), EVA-02 (Fang et al., 2024), ViTamin (Chen et al., 2024) and variants designed to improve robustness or fine-grained understanding, such as NegCLIP (Yuksekgonul et al., 2023), DetailCLIP (Monsefi et al., 2025), and SuperCLIP (Zhao et al., 2025).

3.2 Results

What'sUp Table 1 shows the performance of various CLIP-based models on the center and off-center subsets of **What'sUp**. Across all models, we observe a consistent and significant drop in accuracy when objects are placed away from the center. This degradation is not limited to a specific architecture or training strategy. It persists across different model families as well as recent variants designed to improve robustness or fine-grained perception. Improvements in scale or resolution do not consistently mitigate this issue. Larger models generally improve overall accuracy, but the gap between center and off-center performance remains substantial.

GRID While **What'sUp** provides a realistic testbed, it also contains natural variations in object size and context. To isolate the effect of object location, we create the **GRID** datasets, where we fix the object size and vary its spatial position within a grid. An ideal object recognition model should be invariant to object location. However, as shown in Figure 3, even when object size is fixed, models consistently perform worse when the object is placed off-center.

These results collectively suggest that contrastive VLMs exhibit a systematic center bias, favoring centrally located content over equally relevant off-center information.

Model	center (↑)	off-center (↑)	center bias (↓)
OpenAI CLIP ViT-B/32	62.9	31.9	31.0
OpenAI CLIP ViT-L/14	94.2	**76.7**	**17.5**
OpenCLIP ViT-B/32	65.7	10.5	55.2
OpenCLIP ViT-L/14	87.6	46.0	41.6
RoBERTaCLIP ViT-B/32	64.8	15.7	49.1
CoCa ViT-B/32	39.0	6.70	32.3
EVA02 ViT-B/16	56.2	13.7	42.5
EVA02 ViT-L/14	88.6	35.1	53.5
EVA02 ViT-L/14 (336)	89.5	33.5	56.0
SigLIP (512)	**97.1**	69.6	27.5
ViTamin-B	66.7	34.5	32.2
ViTamin-L	92.3	57.8	34.5
ViTamin-L (336)	**97.1**	67.1	30.0
NegCLIP ViT-B-32	50.4	22.6	27.8
DetailCLIP	28.6	6.40	22.2
SuperCLIP ViT-B/16	50.5	22.4	28.1
SuperCLIP ViT-L/16	60.9	11.8	49.1

Table 1: Performance of CLIP-based models on the center and off-center subsets of **What'sUp**. Despite strong performance on centrally located objects, all models show significantly lower accuracy on off-center examples, indicating a systematic center bias.

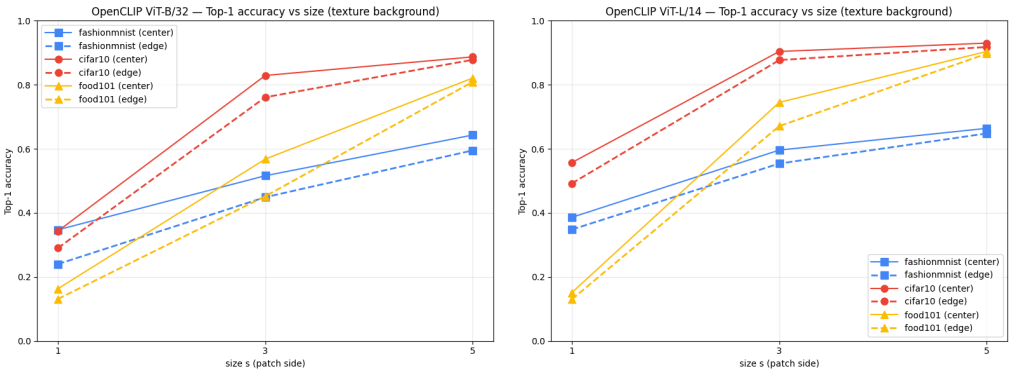

Figure 3: Mean top-1 accuracy vs. object size on a 7×7 **GRID**. Left: OpenCLIP ViT-B/32. Right: OpenCLIP ViT-L/14. Despite improvements with increasing object size, a persistent center–off-center performance gap remains.

4 Why Does CLIP Exhibit Center Bias?

Having established the presence of center bias across datasets and model variants in Section 3.2, we next seek to understand its underlying causes. While center bias is widespread, the exact mechanism may vary across architectures. In this work, we focus on class-token ([CLS])-based CLIP variants, one of the most important and popular designs (e.g., ViT, OpenCLIP models), where the aggregation process is particularly transparent to analyze.

To this end, we analyze CLIP from two complementary perspectives that capture both *what* information is encoded and *how* it is derived. First, we perform **embedding decomposition** to examine how different regions of an image contribute to the final representation and whether central regions disproportionately dominate the embedding. Second, we conduct **attention map analysis** to investigate whether the model inherently allocates more attention to central regions, potentially leading to the observed bias.

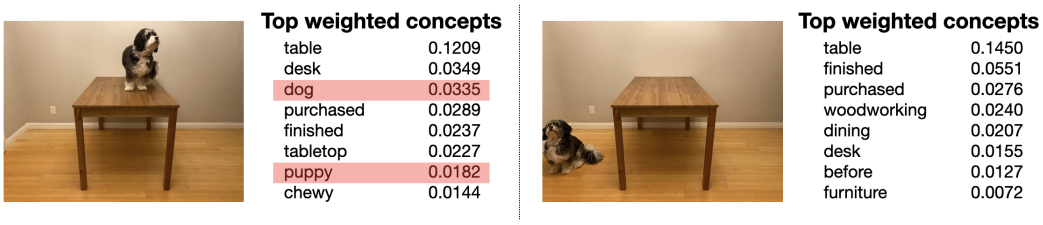

Figure 4: Comparison of top-weighted concepts from CLIP when the dog is placed at the center (left) versus off-center (right). When centered, CLIP assigns meaningful weights to "dog"-related concepts, whereas these concepts disappear when the dog is placed off-center

4.1 Embedding Decomposition

As CLIP relies on a text-aligned vision encoder, we design the following experiment to examine whether its vision embeddings are sufficient for fine-grained understanding. We adopt SpLiCE (Bhalla et al., 2024), an interpretable method that can decompose and explain the information captured by CLIP representations, and apply it to **What'sUp** examples.

Specifically, given an image I and its CLIP vision embedding $\mathbf{x}$, we consider a set of concept keywords $\{t_1, t_2, \cdots, t_n\}$ that may be present in the image. Let $\{\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_n\}$ denote the corresponding CLIP text embeddings for these concept keywords and C denote a matrix with its i-th column being $\mathbf{c}_i$. SpLiCE then finds a set of linear weights $\mathbf{w}$ over the concepts by solving the following optimization problem:

$$\min_{\mathbf{w} \in \mathcal{R}_+^n} ||\mathbf{C}\mathbf{w} - \mathbf{x}||_2^2 + 2\lambda ||\mathbf{w}||_1$$

Figure 4 illustrates an example where a table and a dog appear in both images. However, when the dog is moved from the center of the image (left figure) to the edge (right figure), the concept of "dog" is no longer among the highly weighted concepts. This suggests that CLIP-based embeddings tend to encode a coarse-grained view of the image, emphasizing larger and centrally located objects. Meanwhile, **the concepts of under-representing smaller or off-center elements vanish completely from the model's embedding**. More qualitative results for concept vanishing in other objects and positions pairs are provided in Appendix C.

4.2 Attention Map Analysis

A common approach to extract visual representations from ViTs, including many implementations of CLIP, is to use the representation of the `[CLS]` class token. This token is designed to aggregate information from all image patches through self-attention, and its final-layer embedding is typically used as the image representation.

However, in Figure 5, we find that the attention used to compute the `[CLS]` embedding is excessively concentrated on the central region. While this behavior may suffice for the ordinary caption-matching task, it fails to capture fine-grained details, particularly for objects located near the boundaries. In contrast, other visual tokens still attend to these off-center objects. This suggests that **the relevant information is present in the intermediate representations but is not effectively preserved when aggregated into the `[CLS]` token.**

5 Test-time Mitigation of Center Bias

As we have identified the heart of the problem is the information loss during pooling, we show two simple strategies namely visual prompting and attention redistribution to mitigate center bias for existing contrastive VLMs in this section. Visual prompting requires an external model but is general and applicable to all CLIP variants. Attention redistribution is empirically more consistent, although it is restricted to CLS-based CLIP models.

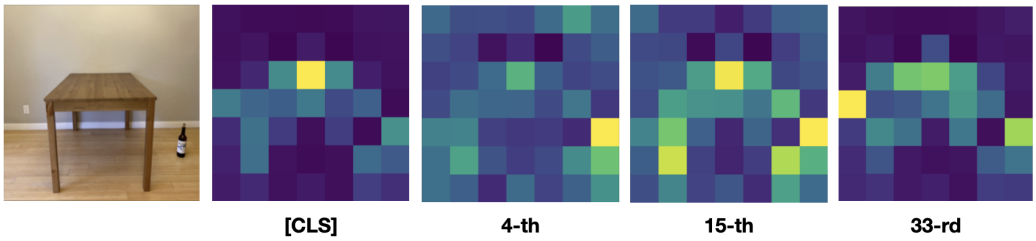

Figure 5: Attention maps for the [CLS] token and other visual tokens in the final layer. Brighter pixels indicate higher attention. While the [CLS] token focuses mainly on the table, patch tokens capture information from other important objects (the bottle and the socket). This suggests that center bias arises not from a lack of available information, but from information loss during pooling.

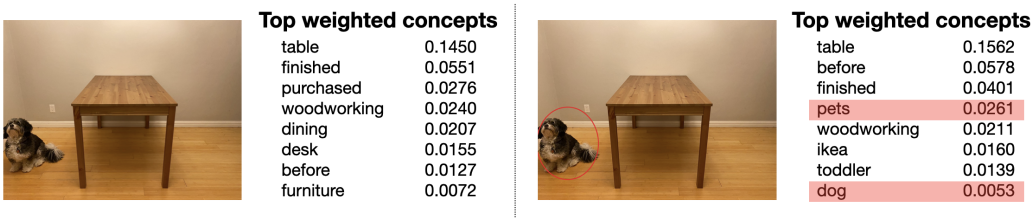

Figure 6: Effect of visual prompting on an off-center object. Without prompting (left), the model fails to assign meaningful "dog"-related concepts. By adding a visual prompt (red circle) around the object (right), relevant concepts such as pets and dog reappear, indicating improved attention to the off-center region.

5.1 Visual Prompting

Prior work has shown that VLMs can be influenced by simple visual cues. Shtedritski et al. (2023) show that model attention can be redirected by drawing a red circle around objects of interest, effectively acting as a visual prompt. Motivated by this observation, we adopt a similar strategy to guide model attention. We use GroundingDINO (Liu et al., 2024) to automatically detect any objects in the image, and then overlay red bounding boxes around the detected regions. These visual prompts encourage the model to attend to potentially overlooked regions, especially when objects are placed off-center. As shown in Figure 6, adding a simple red circle around the object causes "dog"-related concepts to reappear in the SpLiCE decomposition, even when the object is placed away from the center.

5.2 Attention Redistribution

Given our observation in Section 4, a natural way to mitigate center bias is to intervene on the final attention used to form the [CLS] representation. While the [CLS] token assigns disproportionately high attention on the central object, other visual tokens often attend to a broader set of relevant regions, including off-center objects. Motivated by this, we redistribute the final-layer attention of the [CLS] token by suppressing its self-attention and renormalizing the remaining mass over patch tokens. Let $A \in \mathbb{R}^{(N+1)\times(N+1)}$ denote the final-layer attention matrix, where index 0 corresponds to the [CLS] token and N is the number of patches. We set

$$\tilde{A}_{0,0} = 0,$$

and renormalize the rest of the [CLS] row:

$$\tilde{A}_{0,j} = \frac{A_{0,j}}{\sum_{k=1}^{N} A_{0,k}}, \quad j = 1, \ldots, N.$$

Model	center (↑)	off-center (↑)	center bias (↓)	improv. off-center (↑)
Visual Prompting				
OpenAI CLIP ViT-B/32	54.2	21.8	32.4	-10.1
OpenAI CLIP ViT-L/14	87.6	76.7	**10.9**	0.00
OpenCLIP ViT-B/32	59.0	23.3	35.7	+12.8
OpenCLIP ViT-L/14	81.0	56.5	24.5	+10.5
RoBERTaCLIP ViT-B/32	52.3	21.7	30.6	+6.00
CoCa ViT-B/32	23.8	5.10	18.7	-1.60
EVA02 ViT-B/16	65.7	33.8	31.9	+20.1
EVA02 ViT-L/14	86.6	69.0	17.6	+33.9
EVA02 ViT-L/14 (336)	91.4	73.5	17.9	+40.0
SigLIP (512)	88.5	70.9	17.6	+1.30
ViTamin-B	56.1	38.3	17.8	+3.80
ViTamin-L	79.0	56.8	22.2	-1.00
ViTamin-L (336)	88.5	70.9	17.6	+3.80
NegCLIP ViT-B-32	37.1	14.1	23.0	-8.50
DetailCLIP	20.0	10.9	9.10	+4.50
SuperCLIP ViT-B/16	52.4	30.0	22.4	+7.60
SuperCLIP ViT-L/16	50.5	20.1	30.4	+8.30
mean performance	63.2	40.8	22.4	+7.71
Attention Redistribution				
OpenAI CLIP ViT-B/32	76.2	49.2	27.0	+17.3
OpenAI CLIP ViT-L/14	**94.3**	**77.3**	17.0	+0.60
OpenCLIP ViT-B/32	68.6	25.2	43.4	+14.7
OpenCLIP ViT-L/14	89.5	48.9	40.6	+2.90
mean performance	82.2	50.2	32.0	+8.90

Table 2: Performance comparison of visual prompting (VP) and attention redistribution (AR) across various CLIP variants on **What'sUp**. VP generally improves off-center performance but exhibits inconsistent effects across models. In contrast, AR improves off-center performance and overall accuracy more consistently.

This modification preserves the relative importance among visual tokens while forcing `[CLS]` to rely more on patch-level evidence, which in turn reduces the dominance of the center region. We apply attention redistribution to CLS-based CLIP implementations available in OpenCLIP.

An alternative idea is to directly aggregate representations from other patch tokens, for example by taking the mean representation. However, this approach does not work, as the projection layer in CLIP is trained specifically on the `[CLS]` representation and does not generalize well to other aggregation schemes.

5.3 Results

Tables 2 and 3 present the results of visual prompting and attention redistribution on **What'sUp** and **GRID**. Both methods yield net improvements, achieving average gains of 7.71 and 8.90 percentage points in off-center performance, respectively, compared to their vanilla counterparts in the real-world **What'sUp** dataset.

However, the two approaches exhibit different trade-offs. Visual prompting is applicable to all CLIP variants, but on both datasets, despite its ability to reduce center bias, visual prompting sometimes degrades overall accuracy. In contrast, attention redistribution consistently improves off-center performance, while maintaining or improving overall accuracy, but only suitable for class-token based CLIP variants.

These results support our hypothesis that center bias stems from the feature aggregation mechanism, and that redistributing attention toward patch tokens better highlights off-center information. Nevertheless, as discussed in Section 5.1, visual prompting remains a flexible, model-agnostic approach that can be easily applied to any existing architectures.

Model	center (↑)	off-center (↑)	center bias (↓)
OpenAI CLIP ViT-B/32	19.1	**15.7**	**3.40**
+ VP	17.7	13.2	4.50
+ AR	**19.2**	**15.7**	3.50
OpenAI CLIP ViT-L/14	30.5	26.0	**4.50**
+ VP	30.4	25.2	5.2
+ AR	**31.2**	**26.7**	**4.50**
OpenCLIP ViT-B/32	17.2	15.1	**2.07**
+ VP	15.7	12.7	3.03
+ AR	**18.6**	**16.5**	**2.07**
OpenCLIP ViT-L/14	35.2	31.4	3.83
+ VP	34.4	27.9	6.50
+ AR	**35.3**	**31.7**	**3.53**
mean model	25.5	22.1	3.45
+VP	24.6	19.8	4.80
+AR	**26.0**	**22.7**	**3.43**

Table 3: Mean performance across CIFAR10, FashionMNIST, and Food101 comparing visual prompting (VP) and attention redistribution (AR) on the synthetic 7×7 **GRID** with $s = 1$. AR improves both off-center performance and overall accuracy.

6 Conclusion

In this paper, we identify a systematic center bias in the CLIP family, where models excessively focus on centrally located content while overlooking off-center objects. Through experiments across multiple datasets and model variants, we show that this bias leads to consistent performance degradation.

We further analyze this phenomenon for the representative class token-based CLIP from both representation and attention perspectives, revealing that relevant information from off-center regions is not effectively preserved due to information loss during pooling. To alleviate this issue, we propose simple test-time strategies, including visual prompting and attention redistribution, which encourage the model to attend to underrepresented regions. Experimental results demonstrate that these approaches can alleviate center bias without modifying model parameters.

Overall, our findings highlight a potential root cause of perceptual blindness in VLMs and motivate a broader re-examination of representation to enable more reliable and spatially aware visual understanding.

7 Limitations and Future Work

In this work, we focus on revealing and understanding center bias in existing CLIP models and propose simple test-time intervention techniques to alleviate it. While our approaches are relatively simple, we believe the findings could inspire future work to explore more sophisticated test-time strategies for mitigating center bias.

Another promising next step is to prevent the emergence of such biases during the contrastive pretraining stage itself, for example by incorporating stronger augmentations or designing position-invariant objectives.

Finally, our analysis reveals that center bias is widespread across CLIP variants and not unique to class token-based architectures. Understanding how center bias emerges in these alternative architectures is an important direction.

Acknowledgement

This research is partially supported by the National Artificial Intelligence Research Resource Pilot (NAIRR) program (No. NAIRR260216) and Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program (No. CIS251227). Portions of this research were conducted with the advanced computing resources provided by Texas A&M High Performance Research Computing.

References

Dyah Adila, Changho Shin, Linrong Cai, and Frederic Sala. Zero-shot robustification of zero-shot models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=fCeUoDr9Tq.

Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio P. Calmon, and Himabindu Lakkaraju. Interpreting clip with sparse linear concept embeddings (splice). In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 84298–84328. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/996bef37d8a638f37bdfcac2789e835d-Paper-Conference.pdf.

Markus Bindemann. Scene and screen center bias early eye movements in scene viewing. *Vision research*, 50(23):2577–2587, 2010.

Ali Borji, Dicky N Sihite, and Laurent Itti. Quantifying the relative influence of photographer bias and viewing strategy on scene viewing. *Journal of Vision*, 11(11):166, 2011.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

Jieneng Chen, Qihang Yu, Xiaohui Shen, Alan Yuille, and Liang-Chieh Chen. Vitamin: Designing scalable vision models in the vision-language era. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

Sri H Dumpala, Aman Jaiswal, Chandramouli Sastry, Evangelos Milios, Sageev Oore, and Hassan Sajjad. Sugarcrepe++ dataset: Vision-language model sensitivity to semantic and lexical alterations. *Advances in Neural Information Processing Systems*, 37:17972–18018, 2024.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, pp. 105171, 2024.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36:31096–31116, 2023.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language*

Processing, pp. 9161–9175, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.568. URL https://aclanthology.org/2023.emnlp-main.568/.

Darina Koishigarina, Arnas Uselis, and Seong Joon Oh. CLIP behaves like a bag-of-words model cross-modally but not uni-modally. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=DldwXCCP25.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. In *Technical Report, Computer Science Department, University of Toronto*, 2009.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, 2024.

Amin Karimi Monsefi, Kishore Prakash Sailaja, Ali Alilooee, Ser-Nam Lim, and Rajiv Ramnath. DetailCLIP: Detail-oriented CLIP for fine-grained tasks. In *Scaling Self-Improving Foundation Models without Human Supervision*, 2025. URL https://openreview.net/forum?id=80AuuQK4UQ.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. In *Proceedings of the Asian Conference on Computer Vision*, pp. 18–34, 2024.

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11987–11997, October 2023.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9568–9578, 2024.

Po-He Tseng, Ran Carmi, Ian GM Cameron, Douglas P Munoz, and Laurent Itti. Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of vision*, 9(7): 4–4, 2009.

Qizhou Wang, Yong Lin, Yongqiang Chen, Ludwig Schmidt, Bo Han, and Tong Zhang. A sober look at the robustness of clips to spurious features. *Advances in Neural Information Processing Systems*, 37:122484–122523, 2024.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=Ee277P3AYC.

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 *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=KRLUvxh8uaX.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, October 2023.

Zeliang Zhang, Zhuo Liu, Mingqian Feng, and Chenliang Xu. Can CLIP count stars? an empirical study on quantity bias in CLIP. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1081–1086, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.59. URL https://aclanthology.org/2024.findings-emnlp.59/.

Weiheng Zhao, Zilong Huang, Jiashi Feng, and Xinggang Wang. SuperCLIP: CLIP with simple classification supervision. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=EeIEvZlmVg.

Figure 7: Additional examples of concept vanishing under different object positions.

A Dataset Details

What'sUp We use all 418 images from **What'sUp** Subset A, of which 105 are *center* and 313 are *off-center*.

GRID For each of the three datasets (FashionMNIST, CIFAR-10, and Food101), we randomly select 10 classes and sample 100 images per class from the selected classes to form the set of base images. For each base image, we generate two instances: one placed at the center and one placed at a random position along the outer ring. This results in 6,000 examples for each fixed object size s. We consider $s \in \{1, 3, 5\}$ to ensure that the center position is well-defined while preventing the object from exceeding the intended patch boundaries.

B On the definition of *center* in What'sUp

We justify our definitions of *center* and *off-center* by comparing the mean and median distances between the bounding-box centroids and the image center for each position, normalized by the image diagonal length. On average, the "on" position has a significantly smaller distance than the other positions, supporting the validity of our experimental setup.

Position	Mean Distance	Median Distance
Left	0.273	0.288
Right	0.261	0.276
Under	0.192	0.198
On	**0.064**	**0.060**

We also tested the setting where "under" is removed from the off-center cases. The results show no meaningful change, suggesting that "under" behaves similarly to "left" and "right".

Model	Center	Off-center	Center-bias
OpenAI CLIP ViT-B/32	62.9	32.9	30.00
OpenAI CLIP ViT-L/14	94.2	80.4	13.8
OpenCLIP ViT-B/32	65.7	8.0	57.7
OpenCLIP ViT-L/14	87.6	41.9	45.7

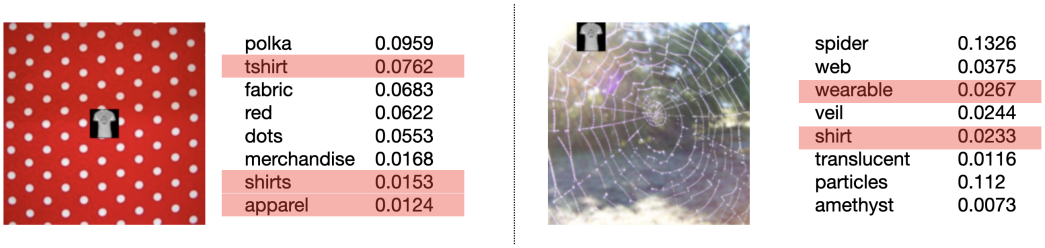

Figure 8: Additional examples of concept vanishing under different object positions.

C More Examples of Concept Vanishing

We provide additional qualitative examples of 'concept vanishing' across varying spatial positions (center vs. left/under) in Figure 7 for **What'sUp**. Consistent with our findings in the main text, the model's representation is heavily biased toward the central object (e.g., armchair, chair), leaving off-center elements under-represented.

The general trend remains the same on GRID. As the object moves from the center to off-center regions, the corresponding concept becomes less represented in the final embedding, and its coefficient and rank in the decomposition analysis decrease (Figure 8).