Pruning the Paradox: How CLIP's Most Informative Heads Enhance Performance While Amplifying Bias

Anonymous ACL submission

Abstract

CLIP is one of the most popular foundational models and is heavily used for many vision-language tasks. However, little is known about the inner workings of CLIP. While recent work has proposed decomposition-based interpretability methods for identifying textual descriptions of attention heads in CLIP, the implications of conceptual consistency in these text labels on interpretability and model performance has not been explored. To bridge this gap, we study the conceptual consistency of text descriptions for attention heads in CLIP-like models. We conduct extensive experiments on six different models from OpenAI and OpenCLIP which vary by size, type of pre-training data and patch size. We propose Concept Consistency Score (CCS), a novel interpretability metric that measures how consistently individual attention heads in CLIP models align with specific concepts. To assign concept labels to heads, we use in-context learning with ChatGPT, guided by a few manually-curated examples, and validate these labels using an LLM-as-a-judge approach. Our soft-pruning experiments reveal that high CCS heads are critical for preserving model performance, as pruning them leads to a significantly larger performance drop than pruning random or low CCS heads. Notably, we find that high CCS heads capture essential concepts and play a key role in out-of-domain detection, concept-specific reasoning, and video-language understanding. Moreover, we prove that high CCS heads learn spurious correlations amplifying social biases. These results position CCS as a powerful interpretability metric exposing the paradox of performance and social biases in CLIP models.

1 Introduction

Large-scale vision-language (VL) models such as CLIP (Radford et al., 2021) have significantly advanced state-of-the-art performance in vision tasks in recent years. Consequently, CLIP has been extensively used as a foundational model for downstream tasks such as video retrieval, image generation, and segmentation (Luo et al., 2022; Liu et al., 2024; Brooks et al., 2023; Esser et al., 2024; Kirillov et al., 2023). This has enabled the construction of compositional models combining CLIP with other foundation models, thereby increasing the functionality of CLIP while also adding complexity to the overall model structure. However, as these models gain prominence in real-world applications, their embedded social biases (Howard et al., 2024; Hall et al., 2023; Seth et al., 2023) have emerged as a critical concern with potentially harmful downstream consequences. Despite the growing body of work documenting these biases, a fundamental question remains unanswered: *what mechanisms within these models' architectures drive both their impressive capabilities and problematic shortcomings?*

Recent interpretability advances (Gandelsman et al.) have made initial progress by decomposing CLIP's image representations into contributions from individual attention heads, identifying text sequences that characterize different heads' semantic roles. However, this approach provides only a partial view into CLIP's inner workings, leaving a critical missing piece: systematic understanding of the visual concepts encoded at the attention head level—and how these concepts underpin both the model's strengths and its social failures.

Our work addresses this critical gap through a novel interpretability framework we call "conceptual consistency". This framework systematically analyzes which visual concepts are learned by individual attention heads and how consistently these concepts are processed throughout the model's architecture. First, we identify interpretable structures within the individual heads of the last four layers of the model using a set of text descriptions. To accomplish this, we employ the TEXTSPAN algorithm (Gandelsman et al.), which helps us find

the most appropriate text descriptions for each head. After identifying these text descriptions, we assign labels to each head representing the common property shared by the descriptions. This labeling process is carried out using in-context learning with ChatGPT. We begin by manually labeling five pairs of text descriptions and their corresponding concept labels, which serve as examples. These examples are then used to prompt ChatGPT to assign labels for the remaining heads.

Leveraging the resulting text descriptions and concept labels of attention heads, we introduce the Concept Consistency Score (CCS), a new interpretability metric that quantifies how strongly individual attention heads in CLIP models align with specific concepts. Using GPT-4o, Gemini and Claude as automatic judges, we compute CCS for each head and classify them into high, moderate, and low categories based on defined thresholds. A key contribution of our work is our targeted soft-pruning experiments which show that heads with high CCS are essential for maintaining model performance; pruning these heads causes a significantly larger performance drop compared to pruning any other heads. We also show that high CCS heads are not only crucial for general vision-language tasks but are especially important for out-of-domain detection and targetted concept-specific reasoning. Additionally, our experiments in video retrieval highlight that high CCS heads are equally vital for temporal and cross-modal understanding. Moreover, we demonstrate that high CCS heads often encode spurious correlations, contributing to social biases in CLIP models. Selective pruning of these heads can reduce such biases without the need for fine-tuning. Together, these results expose a fundamental paradox: while high-CCS heads are indispensable for strong model performance, they are simultaneously key contributors to undesirable biases.

2 Related Work

Early research on interpretability primarily concentrated on convolutional neural networks (CNNs) due to their intricate and opaque decision-making processes (Zeiler and Fergus, 2014; Selvaraju et al., 2017; Simonyan et al., 2014; Fong and Vedaldi, 2017; Hendricks et al., 2016). More recently, the interpretability of Vision Transformers (ViT) has garnered significant attention as these models, unlike CNNs, rely on self-attention mechanisms rather than convolutions. Researchers have focused on task-specific analyses in areas such as image classification, captioning, and object detection to understand how ViTs process and interpret visual information (Dong et al., 2022; Elguendouze et al., 2023; Mannix and Bondell, 2024; Xue et al., 2022; Cornia et al., 2022; Dravid et al., 2023). One of the key metrics used to measure interpretability in ViTs is the attention mechanism itself, which provides insights into how the model distributes focus across different parts of an image when making decisions (Cordonnier et al., 2019; Chefer et al., 2021). This has led to the development of techniques that leverage attention maps to explain ViT predictions. Early work on multimodal interpretability, which involves models that handle both visual and textual inputs, probed tasks such as how different modalities influence model performance (Cao et al., 2020; Madasu and Lal, 2023) and how visual semantics are represented within the model (Hendricks and Nematzadeh, 2021; Lindström et al., 2021). Aflalo et al. (Aflalo et al., 2022) explored interpretability methods for vision-language transformers, examining how these models combine visual and textual information to make joint decisions. Similarly, Stan et al. (Stan et al., 2024) proposed new approaches for interpreting vision-language models, focusing on the interactions between modalities and how these influence model predictions. Our work builds upon and leverages the methods introduced by Gandelsman et al., 2024) to interpret attention heads, neurons, and layers in vision-language models, providing deeper insights into their decision-making processes.

3 Quantifying interpretability in CLIP models

3.1 Preliminaries

In this section, we describe our methodology, starting with the TEXTSPAN (Gandelsman et al.) algorithm and its extension across all attention heads in multiple CLIP models using in-context learning. TEXTSPAN associates each attention head with relevant text descriptions by analyzing the variance in projections between head outputs and candidate text representations. Through iterative projections, it identifies distinct components aligned with different semantic aspects. While effective at linking heads to descriptive text spans, TEXTSPAN does not assign explicit concept labels. In the next section, we detail our method for labeling the concepts

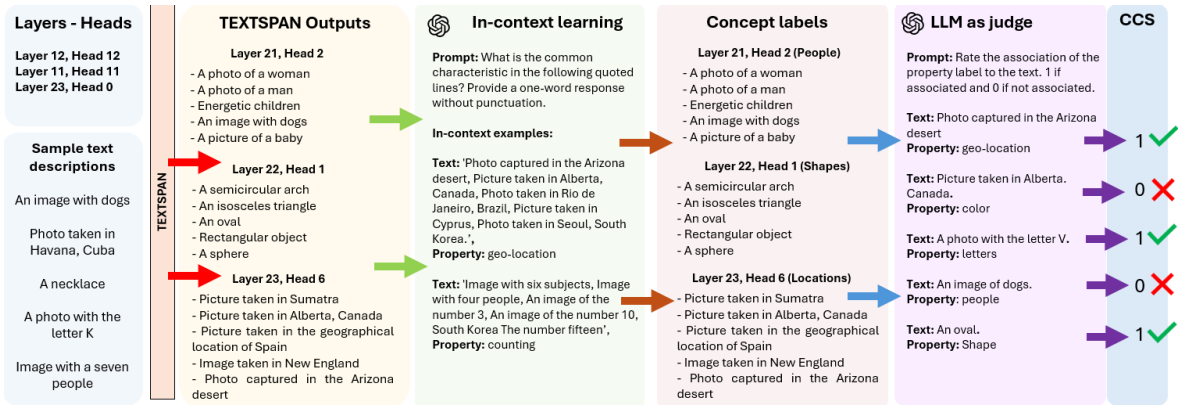


Figure 1: **Figure shows the steps of computing Concept Consistency Score for each head.**

High CCS ($CCS = 5$)	**Moderate CCS** ($CCS = 3$)	**Low CCS** ($CCS \leq 1$)
L23.H11 ("People")	**L23.H0 ("Material")**	**L21.H6 ("Professions")**
Playful siblings	Intrica wood carvingte	Photo taken in the Italian pizzerias
A photo of a young person	Nighttime illumination	thrilling motorsport race
Image with three people	Image with woven fabric design	Urban street fashion
A photo of a woman	Image with shattered glass reflections	An image of a Animal Trainer
A photo of a man	A photo of food	A leg
L22.H10 ("Animals")	**L11.H0 ("Locations")**	**L10.H6 ("Body parts")**
Image showing prairie grouse	Photo taken in Monument Valley	A leg
Image with a donkey	Majestic animal	colorful procession
Image with a penguin	An image of Andorra	Contemplative monochrome portrait
Image with leopard print patterns	An image of Fiji	Graceful wings in motion
detailed reptile close-up	Image showing prairie grouse	Inviting reading nook
L23.H5 ("Nature")	**L11.H11 ("Letters")**	**L9.H2 ("Textures")**
Intertwined tree branches	A photo with the letter J	Photo of a furry animal
Flowing water bodies	A photo with the letter K	Closeup of textured synthetic fabric
A meadow	A swirling eddy	Eclectic street scenes
A smoky plume	A photo with the letter C	Serene beach sunset
Blossoming springtime blooms	awe-inspiring sky	Minimalist white backdrop

Table 1: **Examples of high, moderate and low CCS heads.**

Model	Kappa	SC (ρ)	Kendall (τ)
ViT-B-32-OpenAI	0.821	0.737	0.781
ViT-B-16-LAION	0.813	0.773	0.737
ViT-L-14-OpenAI	0.827	0.751	0.758

Table 2: **Results between human judgment and LLM judgment on CCS labelling**. SC denotes Spearman's correlation.

learned by individual CLIP heads.

3.2 Concept Consistency Score (CCS)

We introduce the Concept Consistency Score (CCS) as a systematic metric for analyzing the concepts (properties) learned by transformer layers and attention heads in CLIP-like models. This score quantifies the alignment between the textual representations produced by a given head and an assigned concept label. Figure 1 illustrates our approach, with the following sections detailing each step in computing CCS.

3.2.1 Extracting Text Representations

From each layer and attention head of the CLIP model, we obtain a set of five textual outputs, denoted as $\{T_1, T_2, T_3, T_4, T_5\}$, referred to as TEXTSPANS. These outputs serve as a textual approximation of the concepts encoded by the head.

3.2.2 Assigning Concept Labels

Using in-context learning with ChatGPT, we analyze the set of five TEXTSPAN outputs and infer a concept label C_h that best represents the dominant concept captured by the attention head h. This ensures that the label is data-driven and reflects the most salient pattern learned by the head.

3.2.3 Evaluating Concept Consistency

To assess the consistency of a head with respect to its assigned concept label, we employ three state-of-the-art foundational models, GPT-4o, Gemini 1.5 pro and Claude Sonnet as external evaluators. For each TEXTSPAN T_i associated with head h, GPT-

4o determines whether it aligns with the assigned concept C_h. The Concept Consistency Score (CCS) for head h is then computed as:

$$\text{CCS}(h) = \sum_{i=1}^{5} \mathbb{1}[T_i \text{ aligns with } C_h]$$

where $\mathbb{1}[\cdot]$ is an indicator function that returns 1 if T_i to be consistent with C_h, and 0 otherwise. To ensure a high standard of reliability, we define consistency strictly—only if all three LLM judges independently rate T_i as consistent with C_h. This requirement for unanimous agreement minimizes the influence of individual model biases or variability in judgment (Liu and Zhang, 2025), thereby enhancing the robustness and trustworthiness of the overall concept consistency score.

We define $CCS@K$ as the fraction of attention heads in a CLIP model that have a Concept Consistency Score (CCS) of K. This metric provides a global measure of how many heads strongly encode interpretable concepts. A higher $CCS@K$ value indicates that a greater proportion of heads exhibit strong alignment with a single semantic property. Mathematically, $CCS@K$ is defined as:

$$CCS@K = \frac{1}{H} \sum_{h=1}^{H} \mathbb{1}[\text{CCS}(h) = K]$$

where H is the total number of attention heads in the model, $\text{CCS}(h)$ is the Concept Consistency Score of head h, $\mathbb{1}[\cdot]$ is an indicator function that returns 1 if $\text{CCS}(h) = K$, and 0 otherwise. This metric helps assess the overall interpretability of the model by quantifying the proportion of heads that consistently capture well-defined concepts. Table 1 shows the examples of heads with different CCS scores.

Next, we categorize each attention head based on its Concept Consistency Score (CCS) into three levels: high, moderate, and low. A head is considered to have a high CCS if all of its associated text descriptions align with the labeled concept, indicating that the head is highly specialized and likely encodes features relevant to that concept. Moderate CCS heads exhibit partial alignment, with three out of five text descriptions matching the concept label, suggesting that they capture the concept to some extent but not exclusively. In contrast, low CCS heads have zero or only one matching description, implying minimal relevance and indicating that these heads are largely unrelated to the given concept. This categorization provides insight into the degree of concept selectivity exhibited by individual attention heads. Table 1 shows examples of different types of CCS heads.

3.3 Evaluating LLM Judgment Alignment with Human Annotations

In the previous section, we introduced the Concept Consistency Score (CCS), computed using three LLM judges as an external evaluator. This raises an important question: *Are LLM evaluations reliable and aligned with human assessments?* To investigate this, we conducted a human evaluation study comparing LLM-generated judgments with human annotations. We selected 100 TEXTSPAN descriptions from three different models, along with their assigned concept labels, and asked one of the authors to manually assess the semantic alignment between each span and its corresponding label.

Table 2 reports the agreement metrics between human and LLM evaluations, including Cohen's Kappa, Spearman's ρ, and Kendall's τ. The Kappa values exceed 0.8, indicating extremely substantial agreement, while the correlation scores consistently surpass 0.7, confirming strong alignment. These results validate the use of LLMs as reliable evaluators in concept consistency analysis. The high agreement with human judgments suggests that LLMs can effectively assess semantic coherence, offering a scalable alternative to manual annotation. In the next section, we introduce the tasks and datasets used in our experiments.

3.4 Experimental Setting

3.4.1 Tasks

Image classification: CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100 (Krizhevsky et al., 2009), Food-101 (Bossard et al., 2014), Country-211 (Radford et al., 2021) and Oxford-pets (Parkhi et al., 2012).
Out-of-domain classification: Imagenet-A (Hendrycks et al., 2021b) and Imagenet-R (Hendrycks et al., 2021a).
Video retrieval: MSRVTT (Xu et al., 2016), MSVD (Chen and Dolan, 2011), DiDeMo (Anne Hendricks et al., 2017).
Bias: FairFace (Karkkainen and Joo, 2021), SocialCounterFactuals (Howard et al., 2024).

Model	CIFAR-10			CIFAR-100			FOOD-101		
	Original	High CCS	Low CCS	Original	High CCS	Low CCS	Original	High CCS	Low CCS
ViT-B-32-OpenAI	75.68	71.31	73.61	65.08	56.07	62.39	84.01	73.42	82.12
ViT-B-32-datacomp	72.07	70.50	70.43	54.95	53.14	53.72	41.66	38.13	40.77
ViT-B-16-OpenAI	78.10	63.93	76.44	68.22	51.70	65.38	88.73	76.35	87.36
ViT-B-16-LAION	82.82	78.91	75.38	76.92	65.55	72.51	86.63	67.54	81.4
ViT-L-14-OpenAI	86.94	86.29	85.97	78.28	75.66	77.55	93.07	90.75	92.79
ViT-L-14-LAION	88.29	86.48	88.19	83.37	80.07	83.25	91.02	86.45	90.35

Table 3: **Accuracy comparison of various CLIP models on CIFAR-10, CIFAR-100 and FOOD-101 datasets. The values represent original accuracy, performance after pruning high-CCS heads, and performance after pruning low-CCS heads.**

3.4.2 Models

For experiments we use the following six foundational image-text models: ViT-B-32, ViT-B-16 and ViT-L-14 pretrained from OpenAI-400M (Radford et al., 2021) and LAION2B (Schuhmann et al., 2022). Next, we discuss in detail the results from the experiments.

4 Results and Discussion

4.1 Interpretable CLIP Models: The Role of CCS.

In this section we examine the role of the Concept Consistency Score (CCS) in revealing CLIP's decision-making process, focusing on the question: *How does CCS provide deeper insights into the functional role of individual attention heads in influencing downstream tasks?* To explore this, we perform a soft-pruning analysis by zeroing out attention weights of heads with extreme CCS values—specifically, high CCS (CCS = 5) and low CCS (CCS $\leq$ 1). This approach disables selected heads without modifying the model architecture. As shown in Table 3, pruning high-CCS heads consistently causes significant drops in zero-shot classification performance across CIFAR-10, CIFAR-100 and FOOD-101 while pruning low-CCS heads has a minimal effect. This performance gap demonstrates that CCS effectively identifies heads encoding critical, concept-aligned information, making it a reliable tool for interpreting CLIP's internal decision-making mechanisms.

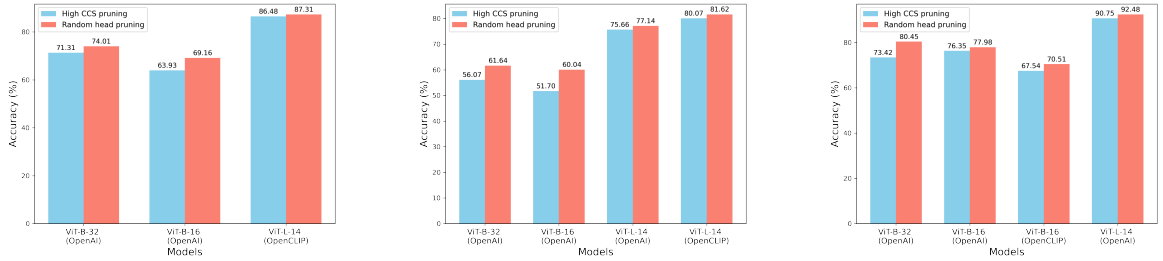
We further observe notable variations in pruning sensitivity across model architectures. ViT-B-16 models suffer the most from high-CCS head pruning, implying a reliance on a smaller number of specialized heads. In contrast, ViT-L-14 models show greater resilience, suggesting more distributed representations. Among smaller models, OpenAI-trained models experience larger performance drops than OpenCLIP models when high-CCS heads are pruned. However, in larger models like ViT-L-14, OpenCLIP variants show a slightly higher degradation. These patterns reveal that CCS not only identifies functionally important heads but also captures model-specific and training-specific differences in how conceptual knowledge is organized and utilized within CLIP architectures.

4.2 High CCS vs random heads pruning

In the previous section, we showed that attention heads with high Concept Consistency Scores (CCS) are crucial to CLIP's performance. To validate whether these heads are truly more important than others, we perform a controlled comparison against random pruning. Specifically, we randomly prune the same number of attention heads—excluding high-CCS heads—and repeat this across five seeds, averaging the results. As illustrated in Figure 2, pruning high-CCS heads consistently causes a significantly larger drop in zero-shot accuracy compared to random pruning across datasets and model variants. In contrast, random pruning results in only minor performance degradation, highlighting the functional importance of high-CCS heads. Interestingly, we also find that larger CLIP models show a smaller performance gap between high-CCS and random pruning, suggesting that larger architectures may be more robust due to greater redundancy or more distributed representations. These findings support CCS as a reliable and interpretable metric for identifying concept-relevant heads and offer deeper insights into how CLIP organizes conceptual information.

4.3 High CCS heads are crucial for out-of-domain (OOD) detection

While our earlier experiments primarily focused on in-domain datasets such as CIFAR-10 and CIFAR-100 to validate the Concept Consistency Score



(a) CIFAR-10 (b) CIFAR-100 (c) Food-101

Figure 2: **Zero-shot performance comparison for CIFAR-10, CIFAR-100, and Food-101 datasets under different pruning strategies.** For random pruning, results are averaged across five runs.

Model	Country-211			Oxford-pets			ImageNet-A			ImageNet-R		
	Original	High CCS	Low CCS	Original	High CCS	Low CCS	Original	High CCS	Low CCS	Original	High CCS	Low CCS
ViT-B-32-OpenAI	17.16	11.46	16.3	50.07	46.66	48.96	31.49	20.24	28.72	69.09	54.47	64.45
ViT-B-32-datacomp	4.43	4.37	4.37	26.48	25.98	25.33	4.96	4.59	4.65	34.06	31.6	32.47
ViT-B-16-OpenAI	22.81	10.72	21.79	52.72	49.12	51.89	49.85	25.49	47.27	77.37	55.52	74.84
ViT-B-16-LAION	20.45	7.49	16.87	65.79	48.48	49.81	37.97	25.27	27.44	80.56	66.32	71.73
ViT-L-14-OpenAI	31.91	23.21	30.63	61.79	62.04	62.08	70.4	68.15	69.2	87.87	86.56	86.97
ViT-L-14-LAION	26.41	16.38	25.66	54.1	56.12	57.16	53.8	42.44	52.93	87.12	82.22	86.94

Table 4: **Accuracy comparison of various CLIP models on Country-211, Oxford-pets, ImageNet-A and ImageNet-R datasets. The values represent original accuracy, performance after pruning high-CCS heads, and performance after pruning low-CCS heads.**

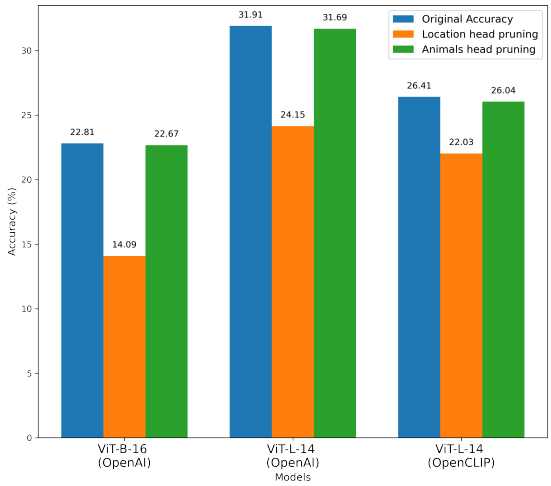


Figure 3: **Zero-shot results on Country-211 (location) dataset.**

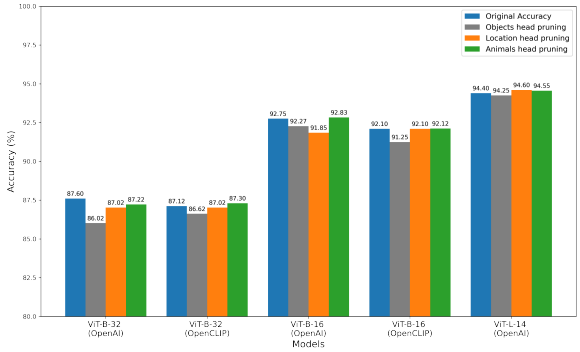
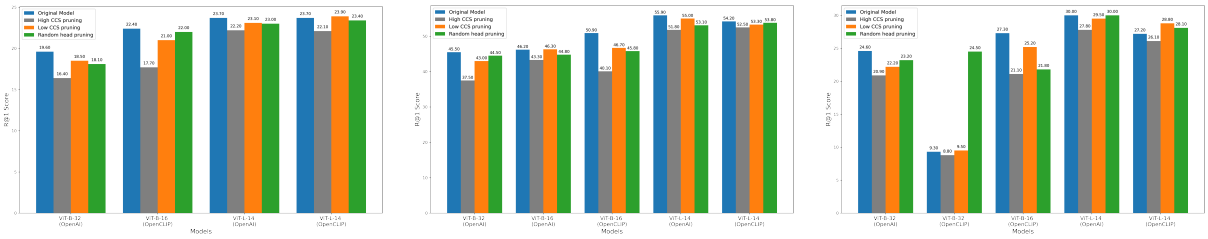


Figure 4: **Zero-shot results on CIFAR-10 (Objects) dataset.**

(CCS), understanding model behavior under out-of-domain (OOD) conditions is a critical step toward evaluating models' robustness. Table 4 demonstrates the results on ImageNet-A and ImageNet-R datasets respectively. From the table, we observe that pruning heads with high CCS scores leads to a substantial degradation in model performance, underscoring the critical role these heads play in the model's decision-making process. Notably, the ViT-B-16-OpenAI model exhibits the most pronounced drop in performance upon pruning high CCS heads, suggesting that this model relies heavily on a smaller set of concept-specific heads for robust feature representation consistent with the observations previously. These results demonstrate that CCS is a powerful metric for identifying attention heads that encode essential, generalizable concepts in CLIP models.



(a) MSRVTT (b) MSVD (c) DIDEMO

Figure 5: **Zero-shot performance comparison of unpruned (original) model, pruning high CSS, low CSS and random heads on video retrieval task.**

4.4 High CCS heads are crucial for concept-specific tasks.

To investigate the functional role of high Concept Consistency Score (CCS) heads, we conduct concept-specific pruning experiments. In these experiments, we prune heads with high CCS scores corresponding to a target concept (e.g., locations) and evaluate the model's performance on tasks aligned with that concept, such as location classification. In contrast, we also prune heads associated with unrelated concepts (e.g., animals) and assess the resulting impact on task performance. Our results indicate that pruning high CCS heads leads to a significant drop in task performance, validating that these heads encode essential concept-relevant information. For instance, in the ViT-B-16 model, pruning location heads results in a substantial decrease in location classification accuracy from 22.81% to 14.09%, as shown in Figure 3. Conversely, pruning heads corresponding to unrelated concepts has little effect on performance, demonstrating the concept-specific nature of high CCS heads, as illustrated in Figure 4.

In more general classification tasks, object-related heads consistently exhibit a greater impact on performance than location or color heads. For example, in the ViT-B-32 model, pruning object-related heads leads to a more noticeable accuracy drop (from 87.6% to 86.02%) compared to pruning location or color heads, which result in smaller reductions (87.02% and 87.22%, respectively). This underscores the greater importance of object-related features in vision tasks. Larger models, such as ViT-L-14, demonstrate a more robust performance to pruning, with smaller accuracy drops when pruning concept-specific heads, suggesting that these models employ more distributed and redundant representations. For instance, pruning object-related heads in ViT-L-14 reduces accuracy only marginally, from 92.1% to 91.25%, with negligible effects from pruning location and color heads. These results not only confirm the effectiveness of CCS as an interpretability tool but also show that high CCS heads are critical for concept-aligned tasks and provide significant insights into how concepts are represented within CLIP-like models.

4.5 Impact of CCS pruning on zero-shot video retrieval.

To further assess the importance of high CCS heads for downstream tasks, we conducted a series of zero-shot video retrieval experiments on three popular datasets: MSRVTT, MSVD, and DIDEMO under different pruning strategies. Figure 5 shows the results of this experiment. Notably, pruning high CCS (Concept Consistency Score) heads consistently leads to a substantial drop in performance across all datasets, demonstrating their critical role in preserving CLIP's retrieval capabilities. For instance, on MSRVTT and MSVD, high CCS pruning significantly underperforms compared to low CCS and random head pruning, which show much milder performance degradation. Interestingly, low CCS and random head pruning maintain performance much closer to the original unpruned model, indicating that not all attention heads contribute equally to model competence. This consistent trend across datasets highlights that heads with high CCS scores are essential for encoding concept-aligned information necessary for accurate zero-shot video retrieval.

4.6 CLIP's high-CCS heads encode features that drive social biases.

Previously, we established that high-CCS heads in CLIP models are crucial for image and video tasks and pruning them leads to significant drop in performance. Now, we investigate if these high CCS heads learn spurious features leading to so-

Model	Race		Gender	
	Original	High CCS	Original	High CCS
ViT-B-32-OpenAI	61.75	60.47	41.24	41.11
ViT-B-32-datacomp	49.2	48.32	21.35	20.97
ViT-B-16-OpenAI	64.61	55.55	40.19	38.21
ViT-B-16-LAION	59.21	56.72	43.55	43.11
ViT-L-14-OpenAI	59.28	59.75	34.7	32.23
ViT-L-14-LAION	61.92	55.59	43.02	39.17

Table 5: **Comparison of original and high-CCS pruning on FairFace dataset for race and gender.**. We used MaxSkew (K=900) as the metric.

Model	Race		Gender	
	Original	High CCS	Original	High CCS
ViT-B-32-OpenAI	3.65	2.43	4.05	1.22
ViT-B-16-OpenAI	2.43	1.22	0.81	2.03
ViT-L-14-OpenAI	2.03	0.81	2.42	1.62

Table 6: **Comparison of original and high-CCS soft-pruning on SocialCounterFactuals dataset for race and gender.**. We used MaxSkew (K=12 for race, K=4 for gender) as the metric.

cial biases. For this, we perform soft pruning experiment on FairFace and SocialCounterFactuals datasets. Here given neutral text prompts of 104 occupations[1], we measure MaxSkew across race and gender in the datasets. Tables 5 and 6 show the results on FairFace and SocialCounterFactuals datasets respectively.

On the FairFace dataset, pruning high-CCS heads consistently reduces the MaxSkew values for both race and gender across all models. For example, in the ViT-B-16-OpenAI model, pruning high-CCS heads drops the race MaxSkew from 64.61 to 55.55 and the gender MaxSkew from 40.19 to 38.21. Similar reductions are observed across all ViT-B and ViT-L variants. These drops, although modest in some cases, indicate a consistent trend: high-CCS heads are contributing disproportionately to skewed model predictions. The effect is even more evident on the SocialCounterfactuals dataset, where MaxSkew values drop substantially upon pruning high-CCS heads. For instance, in ViT-B-32-OpenAI, the gender MaxSkew falls from 4.05 to 1.22, and race MaxSkew from 3.65 to 2.43. Similar reductions occur for other ViT variants, with some pruned models showing more than 50% decrease in bias.

These results reveal a fundamental paradox at the heart of CLIP models: high-CCS heads, while critical for strong performance in tasks such as classification, retrieval, and concept alignment, are also the primary contributors to social bias. This paradox emerges from CLIP's contrastive learning objective, which optimizes alignment between images and their paired text across large, uncurated datasets. In doing so, the model often absorbs and amplifies spurious correlations between visual features and demographic attributes. High-CCS heads, by virtue of their consistent focus on semantically aligned regions, become particularly susceptible to reinforcing these correlations. Pruning these heads leads to a notable reduction in model bias, as shown in our experiments, but also comes at the cost of reduced performance—a clear trade-off between fairness and utility. This performance-bias paradox underscores the complex role of high-CCS heads: they are both enablers of semantic understanding and carriers of learned stereotypes. The CCS metric, in this context, provides a valuable lens for navigating this tension. It not only aids in interpreting model behavior but also offers a lightweight intervention—soft-pruning—that mitigates bias without requiring expensive fine-tuning.

5 Conclusion

In this work, we proposed Concept Consistency Score (CCS), a novel interpretability metric that quantifies how consistently individual attention heads in CLIP-like models align with semantically meaningful concepts. Through extensive soft-pruning experiments, we demonstrated that heads with high CCS are essential for maintaining model performance, as their removal leads to substantial performance drops compared to pruning random or low CCS heads. Our findings further highlight that high CCS heads are not only critical for standard vision-language tasks but also play a central role in out-of-domain detection and concept-specific reasoning. Moreover, experiments on video retrieval tasks reveal that high CCS heads are crucial for capturing temporal and cross-modal relationships, underscoring their broad utility in multimodal understanding. In addition, we demonstrated that high-CCS heads learn spurious correlations leading to social biases and pruning them mitigates that harmful behaviour without the need for further fine-tuning. Thus, CCS provides an wholistic view of interpretability proving the paradox performance vs social biases in CLIP.

[1]List of occupations and prompts can be foudn in Appendix

6 Limitations

In this work, we experimented primarily on CLIP models. Although CCS metric established the fundamental paradox of performance vs social biases we haven't proved for other vision language models. Hence, we leave extending for more vision language models for future work. Another limitation is the use of LLM models for concept labelling and judging which requires robust manual verification to limit any inconsistencies. Hence, scaling our work to much bigger models with more layers and heads can be a limitation.

References

Estelle Aflalo, Meng Du, Shao-Yen Tseng, Yongfei Liu, Chenfei Wu, Nan Duan, and Vasudev Lal. 2022. Vl-interpret: An interactive visualization tool for interpreting vision-language transformers. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 21406–21415.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, pages 446–461. Springer.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402.

Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. 2020. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 565–580. Springer.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791.

David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200.

Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. 2019. On the relationship between self-attention and convolutional layers. *arXiv preprint arXiv:1911.03584*.

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2022. Explaining transformer-based image captioning models: An empirical analysis. *AI Communications*, 35(2):111–129.

Bowen Dong, Pan Zhou, Shuicheng Yan, and Wangmeng Zuo. 2022. Towards class interpretable vision transformer with multi-class-tokens. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 609–622. Springer.

Amil Dravid, Yossi Gandelsman, Alexei A. Efros, and Assaf Shocher. 2023. Rosetta neurons: Mining the common units in a model zoo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1934–1943.

Sofiane Elguendouze, Adel Hafiane, Marcilio CP de Souto, and Anaïs Halftermeyer. 2023. Explainability in image captioning based on the latent space. *Neurocomputing*, 546:126319.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, and 1 others. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*.

Ruth C Fong and Andrea Vedaldi. 2017. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, pages 3429–3437.

Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. Interpreting clip's image representation via text-based decomposition. In *The Twelfth International Conference on Learning Representations*.

Yossi Gandelsman, Alexei A. Efros, and Jacob Steinhardt. 2024. Interpreting the second-order effects of neurons in clip. *Preprint*, arXiv:2406.04341.

Siobhan Mackenzie Hall, Fernanda Gonçalves Abrantes, Hanwen Zhu, Grace Sodunke, Aleksandar Shtedritski, and Hannah Rose Kirk. 2023. Visogender: A dataset for benchmarking gender bias in image-text pronoun resolution. *Advances in Neural Information Processing Systems*, 36:63687–63723.

Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. 2016. Generating visual explanations. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 3–19. Springer.

Lisa Anne Hendricks and Aida Nematzadeh. 2021. Probing image-language transformers for verb understanding. *arXiv preprint arXiv:2106.09141*.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, and 1 others. 2021a. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. 2021b. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15262–15271.

Phillip Howard, Avinash Madasu, Tiep Le, Gustavo Lujan Moreno, Anahita Bhiwandiwalla, and Vasudev Lal. 2024. Socialcounterfactuals: Probing and mitigating intersectional social biases in vision-language models with counterfactual examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11975–11985.

Kimmo Karkkainen and Jungseock Joo. 2021. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1548–1558.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, and 1 others. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026.

Alex Krizhevsky, Geoffrey Hinton, and 1 others. 2009. Learning multiple layers of features from tiny images.

Adam Dahlgren Lindström, Suna Bensch, Johanna Björklund, and Frank Drewes. 2021. Probing multimodal embeddings for linguistic properties: the visual-semantic case. *arXiv preprint arXiv:2102.11115*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Ming Liu and Wensheng Zhang. 2025. Is your video language model a reliable judge? In *The Thirteenth International Conference on Learning Representations*.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2022. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304.

Avinash Madasu and Vasudev Lal. 2023. Is multimodal vision supervision beneficial to language? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2637–2642.

Evelyn Mannix and Howard Bondell. 2024. Scalable and robust transformer decoders for interpretable image classification with foundation models. *arXiv preprint arXiv:2403.04125*.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. 2012. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, and 1 others. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Ashish Seth, Mayur Hemani, and Chirag Agarwal. 2023. Dear: Debiasing vision-language models with additive residuals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6820–6829.

K Simonyan, A Vedaldi, and A Zisserman. 2014. Deep inside convolutional networks: visualising image classification models and saliency maps. In *Proceedings of the International Conference on Learning Representations (ICLR)*. ICLR.

Gabriela Ben Melech Stan, Raanan Yehezkel Rohekar, Yaniv Gurwicz, Matthew Lyle Olson, Anahita Bhiwandiwalla, Estelle Aflalo, Chenfei Wu, Nan Duan, Shao-Yen Tseng, and Vasudev Lal. 2024. Lvlm-intrepret: An interpretability tool for large vision-language models. *arXiv preprint arXiv:2404.03118*.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

Mengqi Xue, Qihan Huang, Haofei Zhang, Lechao Cheng, Jie Song, Minghui Wu, and Mingli Song. 2022. Protopformer: Concentrating on prototypical parts in vision transformers for interpretable image recognition. *arXiv preprint arXiv:2208.10431*.

Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In

Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13, pages 818–833. Springer.

A Concept Consistency Scores (CCS) for CLIP models.

We measure $CCS@K$ for all values of K i.e $K \in [0, 5]$. Table 7 presents the Concept Consistency Score (CCS) distribution across various CLIP models, categorized by architecture size, patch size, and pre-training data. Several noteworthy trends emerge from this analysis. First, models pre-trained on larger and more diverse datasets (e.g., OpenCLIP-LAION2B) tend to exhibit a higher proportion of heads with CCS@5, indicating that a greater number of transformer heads are aligned with semantically meaningful concepts. For instance, the ViT-L-14 model trained on LAION2B shows the highest CCS@5 score of 0.328, suggesting that approximately 32.8% of heads are consistently associated with a single concept, reflecting strong concept alignment in these models.

Second, smaller models such as ViT-B-32 trained on OpenAI-400M demonstrate a significantly lower CCS@5 score (0.167) and a higher proportion of heads with lower CCS values (e.g., CCS@0 = 0.021), indicating weaker alignment of heads to consistent concepts. This observation implies that larger models with richer pre-training data are better at learning concept-specific representations, a key requirement for robust and interpretable multimodal reasoning.

Interestingly, when comparing models with the same architecture but different pre-training corpora, such as ViT-B-32 (OpenAI-400M vs. OpenCLIP-datacomp), we observe a higher CCS@5 score for datacomp (0.229) than OpenAI-400M (0.167), suggesting that dataset composition significantly affects the emergence of interpretable heads.

Moreover, progressive increases in CCS from CCS@0 to CCS@5 show how concept alignment varies within each model. For instance, while ViT-L-14 (OpenCLIP-LAION2B) has a low CCS@0 of 0.016, it steadily increases to a high CCS@5 of 0.328, suggesting that although a few heads are poorly aligned, a substantial fraction are highly consistent in capturing specific concepts.

In summary, these results demonstrate that the CCS metric effectively captures differences in conceptual alignment across models of varying size and pre-training datasets. Models with larger capacities and richer pre-training datasets tend to exhibit higher concept consistency, offering better interpretability and potentially stronger generalization abilities. This analysis underscores the value of CCS as a diagnostic tool for evaluating and comparing the internal conceptual representations learned by CLIP-like models.

Model	Model size	Patch size	Pre-training data	CCS@0	CCS@1	CCS@2	CCS@3	CCS@4	CCS@5
CLIP	B	32	OpenAI-400M	0.021	0.062	0.167	0.271	0.312	0.167
CLIP	B	32	OpenCLIP-datacomp	0.104	0.062	0.208	0.189	0.208	0.229
CLIP	B	16	OpenAI-400M	0.021	0.062	0.125	0.292	0.292	0.208
CLIP	B	16	OpenCLIP-LAION2B	0.062	0.062	0.105	0.25	0.25	0.271
CLIP	L	14	OpenAI-400M	0.062	0.109	0.172	0.204	0.203	0.25
CLIP	L	14	OpenCLIP-LAION2B	0.016	0.031	0.109	0.219	0.297	0.328

Table 7: **Concept Consistency Score (CCS) for CLIP models**.

ViT-B-32-OpenAI L8.H11 (Descriptive), L9.H2 (Objects), L9.H3 (Descriptions), L10.H8 (Locations), L11.H1 (Objects), L11.H5 (Colors), L11.H7 (Objects), L11.H9 (Locations)
ViT-B-32-datacamp L8.H1 (Objects), L8.H3 (Subjects), L8.H10 (Objects), L9.H3 (Subjects), L9.H10 (Objects), L10.H7 (Locations), L10.H11 (Objects), L11.H3 (Colors), L11.H4 (Colors), L11.H9 (Colors), L11.H10 (Objects)
ViT-B-16-OpenAI L8.H5 (Visual), L8.H8 (Visual), L10.H5 (Subjects), L10.H7 (Settings), L11.H0 (Creative), L11.H3 (Settings), L11.H4 (Stylistic), L11.H6 (Locations), L11.H7 (Colors), L11.H11 (Animals)
ViT-B-16-LAION L8.H6 (Descriptions), L8.H7 (Descriptions), L9.H0 (Themes), L9.H1 (Aesthetics), L9.H3 (Descriptive), L10.H5 (Artwork), L10.H10 (Locations), L11.H0 (Locations), L11.H2 (Descriptions), L11.H6 (Locations), L11.H7 (Objects), L11.H8 (Objects), L11.H10 (Colors)
ViT-L-14-OpenAI L20.H2 (Locations), L20.H12 (Descriptions), L21.H0 (Locations), L21.H1 (Locations), L21.H8 (Expressions), L21.H13 (Locations), L21.H15 (Locations), L22.H1 (Objects), L22.H2 (Locations), L22.H5 (Locations), L22.H9 (Subjects), L22.H13 (Animals), L22.H15 (Locations), L23.H4 (Objects), L23.H10 (Locations), L23.H11 (Colors)
ViT-L-14-LAION L20.H4 (Subjects), L20.H14 (Descriptions), L21.H0 (Colors), L21.H1 (Locations), L21.H5 (Descriptive), L21.H9 (Colors), L21.H11 (Locations), L22.H0 (Patterns), L22.H1 (Shapes), L22.H3 (Objects), L22.H5 (Visual), L22.H6 (Animals), L22.H8 (Letters), L22.H10 (Colors), L22.H12 (Landscapes), L22.H13 (Locations), L23.H4 (People), L23.H5 (Nature), L23.H6 (Locations), L23.H8 (Colors), L23.H9 (Descriptive)

Table 8: **Full List of high-CCS heads of all CLIP models.**

ViT-B-32-OpenAI L8.H1 (Artistic), L8.H2 (Objects), L8.H6 (Photography), L8.H9 (Styles), L8.H10 (Perspective), L9.H1 (Subjects), L9.H11 (Settings), L10.H0 (Objects), L10.H3 (Locations), L10.H7 (Locations), L11.H6 (Descriptions), L11.H10 (Locations), L11.H11 (Locations)
ViT-B-32-datacamp L8.H0 (Environments), L8.H7 (Creativity), L9.H6 (Colors), L10.H5 (Art), L10.H6 (Descriptions), L10.H8 (Locations), L10.H9 (Descriptions), L11.H2 (Subjects), L11.H8 (Qualities)
ViT-B-16-OpenAI L8.H1 (Artistic), L8.H2 (Photography), L8.H4 (Styles), L8.H6 (Artwork), L8.H7 (Photography), L8.H9 (Light), L9.H4 (Photography), L9.H6 (Artforms), L9.H10 (Elements), L10.H3 (Locations), L10.H8 (Colors), L10.H9 (Artwork), L11.H5 (Objects), L11.H8 (Effects)
ViT-B-16-LAION L8.H0 (Locations), L8.H8 (Text), L8.H9 (Photography), L9.H7 (Artistic), L9.H8 (Settings), L9.H11 (Descriptions), L10.H2 (Nature), L10.H3 (Location), L10.H7 (Expressions), L11.H3 (Settings), L11.H9 (Numbers), L11.H11 (Letters)
ViT-L-14-OpenAI L20.H0 (Locations), L20.H3 (Locations), L20.H7 (Communication), L20.H8 (Vehicles), L20.H10 (Locations), L21.H4 (Photography), L21.H6 (People), L21.H10 (Locations), L22.H3 (Countries), L22.H12 (Professions), L23.H3 (Patterns), L23.H9 (Creativity), L23.H15 (Visual)
ViT-L-14-LAION L20.H0 (Locations), L20.H1 (Locations), L20.H2 (Locations), L20.H8 (Locations), L20.H9 (Locations), L20.H11 (Aesthetics), L20.H15 (Descriptions), L21.H12 (Photography), L21.H14 (Locations), L22.H9 (Activities), L22.H14 (Colors), L22.H15 (Emotions), L23.H0 (Materials), L23.H3 (Settings)

Table 9: **Full List of medium-CCS heads of all CLIP models.**

ViT-B-32-OpenAI L8.H5 (Patterns), L9.H9 (Ambiance), L11.H0 (Diverse), L11.H8 (Word)
ViT-B-32-datacamp L8.H2 (Images), L8.H4 (Varied), L8.H9 (Varied), L9.H4 (Variety), L9.H5 (Professions), L11.H0 (Diverse), L11.H1 (Varied), L11.H11 (Settings)
ViT-B-16-OpenAI L8.H0 (Diversity), L9.H3 (Locations), L10.H6 (Body parts), L11.H2 (Perspective)
ViT-B-16-LAION L8.H4 (Variety), L8.H5 (Varied), L8.H10 (Diverse), L9.H2 (Textures), L10.H6 (Photography), L10.H8 (Traits)
ViT-L-14-OpenAI L20.H1 (Diverse), L20.H4 (Diversity), L20.H6 (Items), L20.H15 (Diverse), L21.H2 (Diversity), L21.H3 (Diverse), L22.H0 (Occupations), L22.H4 (Settings), L22.H6 (Weather), L22.H14 (Items), L23.H5 (Diversity)
ViT-L-14-LAION L20.H13 (Photography), L21.H6 (Professions), L23.H1 (Diverse)

Table 10: **Full List of low-CCS heads of all CLIP models.**

Occupations
biologist, composer, economist, mathematician, model, poet, reporter, zoologist, artist, coach, athlete, audiologist, judge, musician, therapist, banker, ceo, consultant, prisoner, assistant, boxer, commander, librarian, nutritionist, realtor, supervisor, architect, priest, guard, magician, producer, teacher, lawyer, paramedic, researcher, physicist, pediatrician, surveyor, laborer, statistician, dietitian, sailor, tailor, attorney, army, manager, baker, recruiter, clerk, entrepreneur, sheriff, policeman, businessperson, chief, scientist, carpenter, florist, optician, salesperson, umpire, painter, guitarist, broker, pensioner, soldier, astronaut, dj, driver, engineer, cleaner, cook, housekeeper, swimmer, janitor, pilot, mover, handyman, firefighter, accountant, physician, farmer, bricklayer, photographer, surgeon, dentist, pianist, hairdresser, receptionist, waiter, butcher, videographer, cashier, technician, chemist, blacksmith, dancer, doctor, nurse, mechanic, chef, plumber, bartender, pharmacist, electrician

Table 11: **Full list of occupations used for evaluating biases on FairFace and SocialCounterFactuals datasets.**

Prompt	Example
A <occupation>	A biologist
A photo of <occupation>	A photo of biologist
A picture of <occupation>	A picture of biologist
An image of <occupation>	An image of biologist

Table 12: **Prompts used for measuring biases on FairFace and SocialCounterFactuals datasets.**