Quantifying Interpretability in CLIP Models with Concept Consistency

Avinash Madasu[♡] Vasudev Lal[♡] Phillip Howard[♢†]
[♡]Intel Labs [♢]Thoughtworks
{avinash.madasu, vasudev.lal}@intel.com phillip.howard@thoughtworks.com

Abstract

CLIP is a widely used foundational model for vision-language tasks, yet its internal mechanisms remain poorly understood. To address this we introduce Concept Consistency Score (CCS), a new interpretability metric that quantifies how strongly individual attention heads align with coherent visual concepts. Using in-context learning with ChatGPT and an LLM-as-a-judge framework, we assign and validate concept labels across six CLIP models of varying sizes, data types, and patch sizes. Our experiments show that high CCS heads are crucial for maintaining model performance, especially in out-of-domain detection, concept reasoning, and video-language tasks. These findings highlight CCS as an effective tool for interpreting and analyzing CLIP-like models.

1. Introduction

Large-scale vision-language (VL) models like CLIP [29] have driven significant advances in visual understanding tasks and are now widely used in downstream applications such as video retrieval, image generation, and segmentation [4, 13, 21, 24, 25]. Their flexibility has enabled integration with other foundation models, resulting in powerful, compositional systems. However, this increased complexity also amplifies the need for interpretability tools that can reveal how such models make predictions—especially in high-stakes or domain-sensitive applications. Given CLIP's foundational role in VL systems, developing robust interpretability techniques to understand its internal mechanisms is both timely and essential.

In this work, we introduce a novel interpretability metric for CLIP models through the lens of visual concept learning. Building on prior work [15] that decomposes CLIP's image representations into contributions from individual attention heads, we focus on identifying the specific visual concepts that each attention head represents. To achieve this, we analyze the heads in the last four layers of the model using the TEXTSPAN algorithm [15], which retrieves descriptive text spans that best capture each head's behavior. These spans are then grouped and labeled with the help of ChatGPT through in-context learning, where a few manually labeled examples guide the automated assignment of conceptual labels to remaining heads. This process reveals interpretable, semantically meaningful structures across the attention heads, allowing us to view CLIP's inner mechanisms through the lens of learned visual concepts.

Leveraging the resulting text descriptions 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 as an automatic judge, 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 random or low CCS 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 concept-specific reasoning. Additionally, our experiments in video retrieval highlight that high CCS heads are equally vital for temporal and cross-modal understanding, thereby underscoring the broad relevance of CCS in analyzing and interpreting CLIP-like models.

2. Quantifying interpretability in CLIP models

2.1. Preliminaries

In this section, we describe our methodology, starting with the TEXTSPAN [15] 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

[†]Work completed while at Intel Labs.

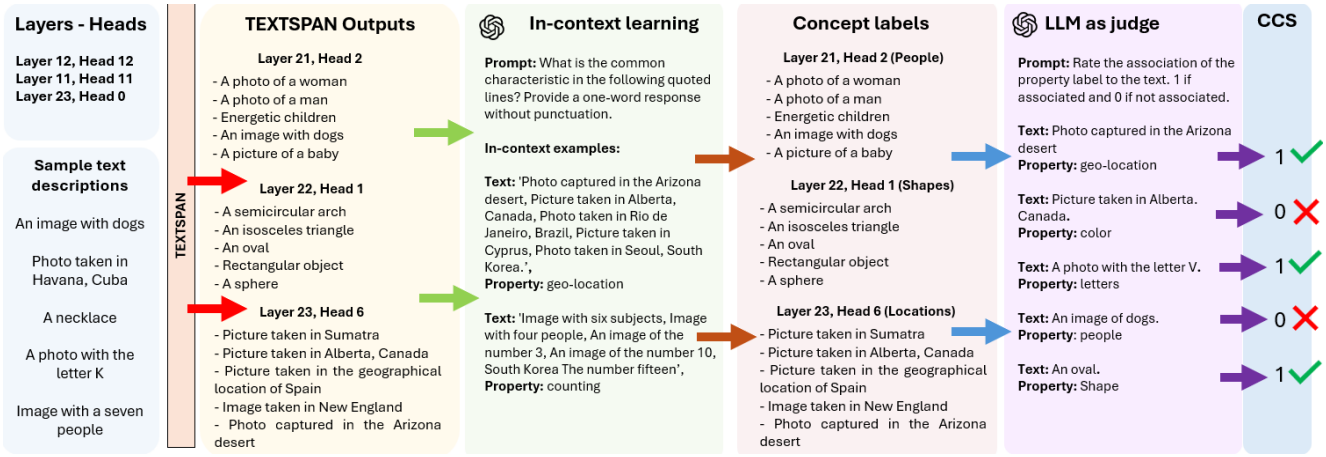


Figure 1. Figure shows the steps of computing Concept Consistency Score for each head.

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

Table 1. **Examples of high, moderate and low CCS heads.** More examples can be found in appendix table 4.

Model	Kappa	SC (ρ)	Kendall (τ)
ViT-B-32-OpenAI	0.757	0.781	0.781
ViT-B-16-LAION	0.676	0.678	0.678
ViT-L-14-OpenAI	0.758	0.758	0.758

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

does not assign explicit concept labels. In the next section, we detail our method for labeling the concepts learned by individual CLIP heads.

2.2. 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.

2.2.1. 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.

2.2.2. 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.

2.2.3. Evaluating Concept Consistency

To assess the consistency of a head with respect to its assigned concept label, we employ a state-of-the-art vision-language model, GPT-4o, as an external evaluator. 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 GPT-4o judges T_i to be consistent with C_h, and 0 otherwise.

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.

2.3. Evaluating LLM Judgment Alignment with Human Annotations

In the previous section, we introduced the Concept Consistency Score (CCS), computed using GPT-4o 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 50 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 0.65, indicating substantial agreement, while the correlation scores consistently surpass 0.65, 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.

2.4. Experimental Setting

2.4.1. Tasks

Image classification: CIFAR-10 [22], CIFAR-100 [22], Food-101 [3], Country-211 [29] and Oxford-pets [28].
Out-of-domain classification: Imagenet-A [20] and Imagenet-R [19].
Video retrieval: MSRVTT [34], MSVD [7], DiDeMo [2].

2.4.2. 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 [29] and LAION2B [30]. Next, we discuss in detail the results from the experiments.

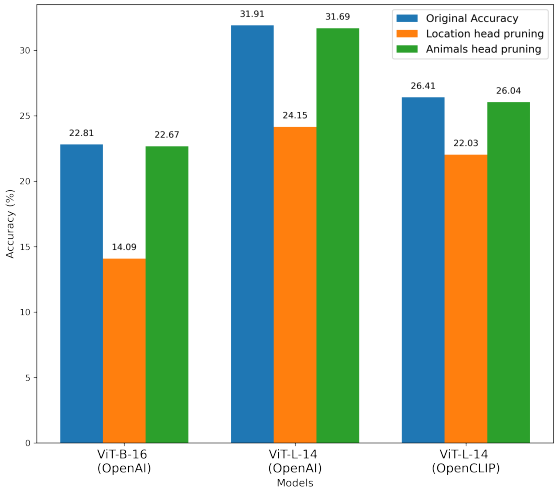


Figure 2. Zero-shot results on Country-211 (location) dataset.

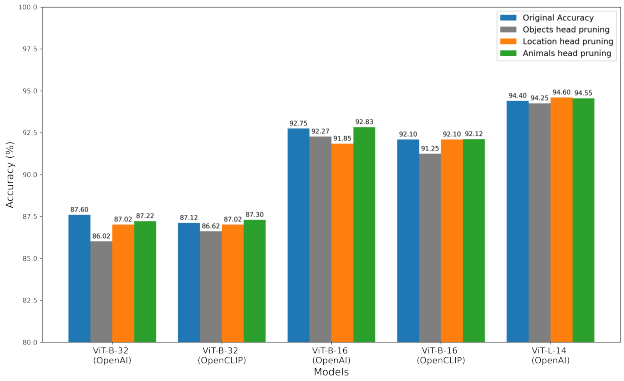


Figure 3. Zero-shot results on CIFAR-10 (Objects) dataset.

3. Results and Discussion

3.1. Interpretable CLIP Models: The Role of CCS.

In this section, we analyze the role of the Concept Consistency Score (CCS) in enhancing CLIP interpretability, 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-

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.

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. We also experiment with pruning equal number of random attention heads as high CCS heads. Results show that high CCS heads leads to significant drop in performance across datasets and variants. Detailed analysis can be found in appendix 5.2.

3.2. 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 2. 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 3.

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. Results for out-of-domain detection and video retrieval can be found in appendix sections 5.3 and 5.4 respectively.

4. 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. Our study shows how CCS is a powerful interpretability metric for identifying key layers and heads in CLIP-like models.

References

[1] Estelle Aflalo, Meng Du, Shao-Yen Tseng, Yongfei Liu, Chenfei Wu, Nan Duan, and Vasudev Lal. 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, 2022. 3

[2] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 3

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 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, 2014. 3

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 1

[5] Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. 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, 2020. 3

[6] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791, 2021. 2

[7] David Chen and William B Dolan. 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, 2011. 3

[8] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. *arXiv preprint arXiv:1911.03584*, 2019. 2

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

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

[11] Amil Dravid, Yossi Gandelsman, Alexei A. Efros, and Assaf Shocher. 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, 2023.

[12] Sofiane Elguendouze, Adel Hafiane, Marcilio CP de Souto, and Anaïs Halftermeyer. Explainability in image captioning based on the latent space. *Neurocomputing*, 546:126319, 2023. 2

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

[14] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, pages 3429–3437, 2017. 2

[15] 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*. 1, 3

[16] Yossi Gandelsman, Alexei A. Efros, and Jacob Steinhardt. Interpreting the second-order effects of neurons in clip, 2024. 3

[17] Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. *arXiv preprint arXiv:2106.09141*, 2021. 3

[18] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. 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, 2016. 2

[19] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. 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, 2021. 3

[20] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15262–15271, 2021. 3

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

[22] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 3

[23] Adam Dahlgren Lindström, Suna Bensch, Johanna Björklund, and Frank Drewes. Probing multimodal embeddings for linguistic properties: the visual-semantic case. *arXiv preprint arXiv:2102.11115*, 2021. 3

[24] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 1

[25] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 1

[26] Avinash Madasu and Vasudev Lal. Is multimodal vision supervision beneficial to language? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2637–2642, 2023. 3

[27] Evelyn Mannix and Howard Bondell. Scalable and robust transformer decoders for interpretable image classification with foundation models. *arXiv preprint arXiv:2403.04125*, 2024. 2

[28] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 3

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 3

[30] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 3

[31] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2

[32] K Simonyan, A Vedaldi, and A Zisserman. Deep inside convolutional networks: visualising image classification models and saliency maps. In *Proceedings of the International Conference on Learning Representations (ICLR)*. ICLR, 2014. 2

[33] 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. Lvlm-intrepret: An interpretability tool for large vision-language models. *arXiv preprint arXiv:2404.03118*, 2024. 3

[34] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 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, 2016. 3

[35] Mengqi Xue, Qihan Huang, Haofei Zhang, Lechao Cheng, Jie Song, Minghui Wu, and Mingli Song. Protopformer: Concentrating on prototypical parts in vision transformers for interpretable image recognition. *arXiv preprint arXiv:2208.10431*, 2022. 2

[36] Matthew D Zeiler and Rob Fergus. 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, 2014. 2

Quantifying Interpretability in CLIP Models with Concept Consistency

Supplementary Material

5. Additional Results

5.1. Concept Consistency Scores (CCS) for CLIP models.

We measure $CCS@K$ for all values of K i.e $K \in [0, 5]$. Table 6 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.

5.2. High CCS vs random heads pruning

To rigorously evaluate the effectiveness of the Concept Consistency Score (CCS) as a measure of interpretability in CLIP models, we compare the impact of pruning heads with high CCS scores against pruning an equal number of randomly selected heads. While earlier results demonstrated that pruning high CCS heads significantly degrades model performance, a critical question remains: *Are these high-CCS heads genuinely more important than other heads?*

To investigate this, we conduct controlled experiments where we randomly prune the same number of attention heads and analyze the corresponding performance drop. The results of these experiments are presented in Figure 5. As shown in the figure, pruning high CCS heads consistently leads to a substantially larger decrease in zero-shot performance compared to random pruning, across different datasets and model variants. These findings empirically validate that CCS effectively identifies heads that are essential for the model's decision-making process, thereby offering a principled mechanism for interpreting the internal workings of CLIP models. Unlike random pruning, which affects heads without regard to their learned properties, CCS-guided pruning systematically targets heads that encode critical concepts, revealing their functional role in model predictions.

Moreover, we observe that larger CLIP models exhibit smaller differences between high CCS and random pruning impacts compared to smaller models, suggesting that larger architectures may possess more redundancy or distributed representations, making them more resilient to head pruning. Overall, these results establish CCS as a reliable and interpretable metric for identifying concept-relevant heads, contributing to a deeper understanding of how CLIP models organize and utilize conceptual knowledge across their layers.

5.3. 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 (CCS), understanding model behavior under out-of-domain (OOD) conditions is a critical step toward evaluating models' robustness and spurious correlations. Table 5 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-

High CCS ($CCS = 5$)	**Moderate CCS** ($CCS = 3$)	**Low CCS** ($CCS \leq 1$)
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 4. **Examples of high, moderate and low CCS heads.**

Model	**Imagenet-A**			**Imagenet-R**		
	Original	**High CCS**	**Low CCS**	**Original**	**High CCS**	**Low CCS**
ViT-B-32-OpenAI	31.49	20.24	28.72	69.09	54.47	64.45
ViT-B-32-datacomp	4.96	4.59	4.65	34.06	31.6	32.47
ViT-B-16-OpenAI	49.85	25.49	47.27	77.37	55.52	74.84
ViT-B-16-LAION	37.97	25.27	27.44	80.56	66.32	71.73
ViT-L-14-OpenAI	70.4	68.15	69.2	87.87	86.56	86.97
ViT-L-14-LAION	53.8	42.44	52.93	87.12	82.22	86.94

Table 5. Accuracy comparison of various CLIP models on ImageNet-A and ImageNet-R. The values represent original accuracy, performance after pruning high-CCS heads, and performance after pruning low-CCS heads.

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 while avoiding spurious correlations.

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

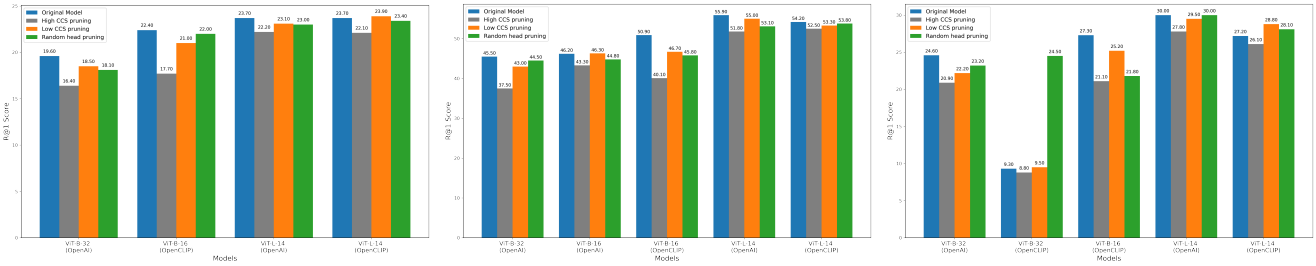
To further assess the importance of high CCS heads for downstream tasks, we conducted a series of zero-shot video retrieval experiments on three popular datasets: MSRVTT, MSVD, and DIDEMO under different pruning strategies. Figure 4 shows the results of this experiment. Notably, pruning high CCS (Concept 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.

6. Related Work

Early research on interpretability primarily concentrated on convolutional neural networks (CNNs) due to their intricate and opaque decision-making processes [14, 18, 31, 32, 36]. 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 [9–12, 27, 35]. 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 [6, 8]. This has led to the development of techniques that leverage attention maps to explain

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 6. **Concept Consistency Score (CCS) for CLIP models**.



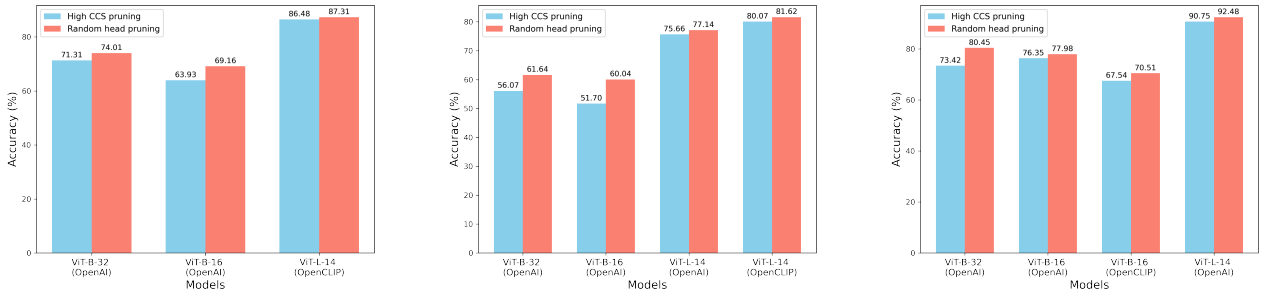
(a) MSRVTT (b) MSVD (c) DIDEMO

Figure 4. **Zero-shot performance comparison of unpruned (original) model, pruning high CSS, low CSS and random heads on video retrieval task.**

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 [5, 26] and how visual semantics are represented within the model [17, 23]. Aflalo et al. [1] explored interpretability methods for vision-language transformers, examining how these models combine visual and textual information to make joint decisions. Similarly, Stan et al. [33] 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. [15, 16] to interpret attention heads, neurons, and layers in vision-language models, providing deeper insights into their decision-making processes.

Model	Country-211			Oxford-pets		
	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
ViT-B-32-datacomp	4.43	4.37	4.37	26.48	25.98	25.33
ViT-B-16-OpenAI	22.81	10.72	21.79	52.72	49.12	51.89
ViT-B-16-LAION	20.45	7.49	16.87	65.79	48.48	49.81
ViT-L-14-OpenAI	31.91	23.21	30.63	61.79	62.04	62.08
ViT-L-14-LAION	26.41	16.38	25.66	54.1	56.12	57.16

Table 7. Accuracy comparison of various CLIP models on Country-211 and Oxford-pets datasets. The values represent original accuracy, performance after pruning high-CCS heads, and performance after pruning low-CCS heads.



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

Figure 5. **Zero-shot performance comparison for CIFAR-10, CIFAR-100, and Food-101 datasets under different pruning strategies.** For random pruning, results are averaged across three runs.