EVOLUTION OF CLIP REPRESENTATIONS: TEXTURE BIAS, HUMAN ALIGNMENT AND NOISE ROBUSTNESS

Pablo Hernández-Cámara
Image Processing Lab
University of Valencia, Spain
pablo.hernandez-camara@uv.es

Jose Manuel Jaén-Lorites
Center for Biomaterials and Tissue Engineering
Universitat Politecnica de Valencia, Spain
jojaelo@doctor.upv.es

Alexandra Gomez-Villa
Computer Vision Center
Universitat Autònoma de Barcelona, Spain
agomezvi@cvc.uab.cat

Jorge Vila-Tomás, Valero Laparra, Jesús Malo
Image Processing Lab
University of Valencia, Spain
{jorge.vila-tomas,valero.laparra,jesus.malo}@uv.es

ABSTRACT

CLIP models demonstrate strong generalization, yet the evolution of their visual representation during training remains poorly understood. We perform an epoch-by-epoch analysis of texture bias, noise robustness, and alignment with human vision across multiple CLIP sizes. We observe a consistent shift from texture-dominated to more shape-sensitive representations as training progresses, though models remain more texture-biased than humans. Despite the improvement in shape-texture bias with further training, alignment with human benchmarks for saliency and mid-low level perception follows an inverse trajectory, reaching its peak early in the training process. As training continues, classification accuracy improves, while perceptual and attentional alignment gradually decline. There is also a trade-off between perceptual alignment and robustness that evolves over the training. These results indicate that different training stages favor different properties, and that selecting models based solely on final accuracy may overlook stages where representations more closely resemble human visual processing.

1 INTRODUCTION

Contrastive Language–Image Pretraining (CLIP) models have set benchmarks in zero-shot learning by aligning visual and linguistic representations Radford et al. (2021); Zhai et al. (2023); Cherti et al. (2023); Chuang et al. (2025). While extensive work has analyzed fully trained models in terms of classification accuracy, noise robustness, and alignment with humans Hendrycks & Dietterich (2019); Liang et al. (2022); Kazemi et al. (2024); Esfandiarpoor et al. (2024); Hernández-Cámara et al. (2024), the evolution of internal visual representations during training remains largely unexplored. In particular, how representational biases develop over time and how this development relates to human perception remains an open question.

Deep neural networks have systematic preferences in visual features. Although both Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) remain substantially more texture-biased than humans, which prioritizes shape for object recognition Landau et al. (1988), CNNs are strongly biased toward texture Geirhos et al. (2018), whereas ViTs and CLIP rely more on shape Geirhos et al. (2021). Existing studies almost exclusively characterize static, final checkpoints of the models, leaving unresolved how these biases emerge and change throughout training. Such a representational evolution is likely to influence both model noise robustness and alignment with human vision.

In this work, we address this gap through a dense epoch-by-epoch analysis of CLIP training across multiple model sizes. We jointly track how multiple properties evolve jointly over training for classification, including texture–shape bias, robustness to noise, and alignment with human perception at different abstraction levels. In particular, we evaluate alignment with low-level image quality judgments, mid-level semantic similarity, and human visual attention patterns. This allows us to

characterize how representational changes during training are reflected in both functional performance and perceptual alignment. Specifically, we find that early training stages exhibit pronounced texture bias together with peak alignment with human perception of low-level distortions and visual saliency. During this same early period, the classification performance of models is more vulnerable to image noise. Interestingly, as training progresses and texture bias decreases, noise robustness improves while alignment with mid-low level perception and attention declines.

These results reveal a shared representational transition linking texture–shape bias, perceptual alignment, attention, and noise sensitivity. Crucially, this pattern is consistent across CLIP sizes (Base, Large, and Huge), indicating a fundamental trade-off between early perceptual alignment and later noise robustness that is intrinsic to CLIP.

2 METHODS

2.1 MODELS AND TRAINING TRAJECTORY

We analized CLIP training dynamics using checkpoints from the OpenCLIP scaling laws project Cherti et al. (2023), trained on the LAION-2B dataset (2B image–text pairs) Schuhmann et al. (2022). The CLIP framework uses a contrastive learning strategy to project images and text descriptions into a shared embedding space. During training, the model is optimized to maximize the cosine similarity between correct image-text pairs while minimizing it for incorrect pairs. This approach allows the visual encoder to learn rich semantic representations that do not rely on fixed class labels, facilitating its zero-shot generalization capabilities. Regarding the architecture, the analyzed models employ a text transformer and a visual transformer (ViT) as textual and visual encoders, respectively. Unlike traditional CNNs, ViTs process images by dividing them into fixed-size patches. In this study, we evaluated the ViT-Base-16, ViT-Large-14, and ViT-Huge-14 variants.

All available checkpoints were evaluated from epoch 0 (random initialization) to the final training epoch, without additional fine-tuning. The results at epoch 0 were averaged over 20 independent random initializations to ensure stability. To assess consistency across model sizes, we replicated all analyses for the three aforementioned CLIP model variants.

2.2 EVALUATION METRICS

To characterize the representational evolution of CLIP, we employed a diverse set of metrics that span semantic performance, structural bias, and human-centric alignment. All metrics were independently computed at each training epoch to track their temporal evolution.

These metrics were selected to provide an evaluation of the trade-offs between functional performance and biological plausibility. By monitoring zero-shot classification, texture-shape bias and noise sensitivity, we can determine whether improvements in accuracy come at the cost of noise robustness or human-like visual processing. Furthermore, incorporating multiple levels of perceptual and saliency alignment allows us to identify specific training stages where representations may be more effective for applications requiring human similarity, rather than just raw categorical precision.

2.2.1 BASELINE PERFORMANCE: ZERO-SHOT CLASSIFICATION ACCURACY

We evaluated classification accuracy as a reference measure of task performance and learning progression. For each checkpoint, we computed the zero-shot accuracy on ImageNet-1K Deng et al. (2009) using the standard CLIP protocol Radford et al. (2021). The image embeddings were compared with the text embeddings generated from class prompts of the form "An image of a {class}", and predictions were obtained via maximum cosine similarity.

2.2.2 FEATURE BIAS ANALYSIS: TEXTURE–SHAPE PREFERENCE

To quantify the preference of the model for texture versus shape features, we evaluated texture–shape bias using the cue-conflict dataset Geirhos et al. (2018). It contains images in which the global shape corresponds to one object category while the local texture corresponds to another.

For each conflict image, we followed the standard Geirhos et al. evaluation procedure Geirhos et al. (2018), in which model outputs over ImageNet classes are mapped to the corresponding shape and texture categories. An image is counted as texture-biased or shape-biased depending on which cue matches the predicted category. The texture bias is then quantified as the proportion of images for which the model prediction aligns with the texture rather than the shape.

2.2.3 Robustness to Noise Perturbations

Noise robustness was assessed by the classification sensitivity to Gaussian noise, which serves as a proxy for environmental stability. We evaluated the zero-shot classification performance on the ImageNet-1K-C Hendrycks & Dietterich (2019) over the five predefined noise amplitudes.

For each training epoch, the accuracy was computed separately for each noise level and subsequently averaged within each epoch. To ensure a consistent comparison across the training trajectory, we report the normalized accuracy drop, defined as the relative decrease from the clean ImageNet-1K accuracy to the mean accuracy across all noise levels. This metric captures the overall sensitivity of the model to noise perturbations and allows us to track how the evolution of internal representations influences the vulnerability of the model to distribution shifts.

2.2.4 Perceptual Alignment

To evaluate the degree to which the internal representations of CLIP mirror human visual processing, we assessed alignment across a hierarchy of perceptual tasks, ranging judgements of low-level image distortions, mid-level similarity, and visual saliency:

Low-level perceptual alignment: Low-level alignment was evaluated as in Kumar et al. (2022); Hernández-Cámara et al. (2025), by measuring the correlation with humans in Image Quality Assessment. Specifically we considered TID2013 Ponomarenko et al. (2015) and KADID-10K Lin et al. (2019) datasets because they are designed to quantify how humans perceive low-level distortions, such as white noise. They consist of image pairs (original and distorted), each accompanied by a Mean Opinion Score, representing the average human judgment of quality. For each epoch, we computed the cosine similarity between the reference embeddings and distorted image pairs. We correlated these similarities with human score using Spearman correlation, providing a measure of the sensitivity of the model to basic visual features.

Mid-level perceptual alignment: To assess sensitivity to more complex visual attributes, we follow the methodology introduced in Hernández-Cámara et al. (2024): we used the NIGHTS perceptual dataset Fu et al. (2023) since (unlike in image quality), it probes mid-level perceptual differences such as changes in object pose, color, and quantity. Each sample presents a triplet: a reference image and two distorted versions, with a ground-truth label indicating which version humans found more similar to the reference. We computed the embedding similarities of the model for both pairs, and then a prediction was counted as correct if the model assigned a higher similarity to the image preferred by humans.

Saliency alignment: We examined the correspondence between model attention and human visual interest using the MIT1003 saliency dataset Judd et al. (2009). Saliency maps represent the spatial distribution of where humans focus their attention when viewing an image. To extract equivalent maps from the model, we employed the attention roll-out method Abnar & Zuidema (2020), which aggregates attention weights across all ViT layers to identify which input patches are prioritized. Alignment is quantified by calculating the correlation between these roll-out maps and human saliency distributions, allowing us to track whether the interest of the model evolves to match human attention during training.

3 Results

To characterize the developmental trajectory of CLIP, we conducted a systematic, epoch-by-epoch evaluation of the models described in section 2.1. Specifically, for every available checkpoint across the ViT-Base, ViT-Large, and ViT-Huge sizes, we independently computed the full suite of metrics described in section 2.2, covering texture bias, noise robustness, and multi-level human alignment.

This comprehensive analysis allows us to observe how representational properties emerge and interact throughout the entire training process.

3.1 Training Dynamics: Texture Bias, Noise Robustness and Alignment

We first confirmed that all models acquire meaningful representations by tracking zero-shot ImageNet-1K accuracy throughout the training (Fig. 1, top-left). Accuracy increased steadily for all model sizes (Base, Large, and Huge), indicating progressive semantic learning. In parallel, we evaluated representational bias using the cue-conflict dataset (Fig. 1, top-center). All models exhibited an early peak in texture bias, followed by a consistent decline over training, revealing a transition from texture-dominated to more shape-based representations.

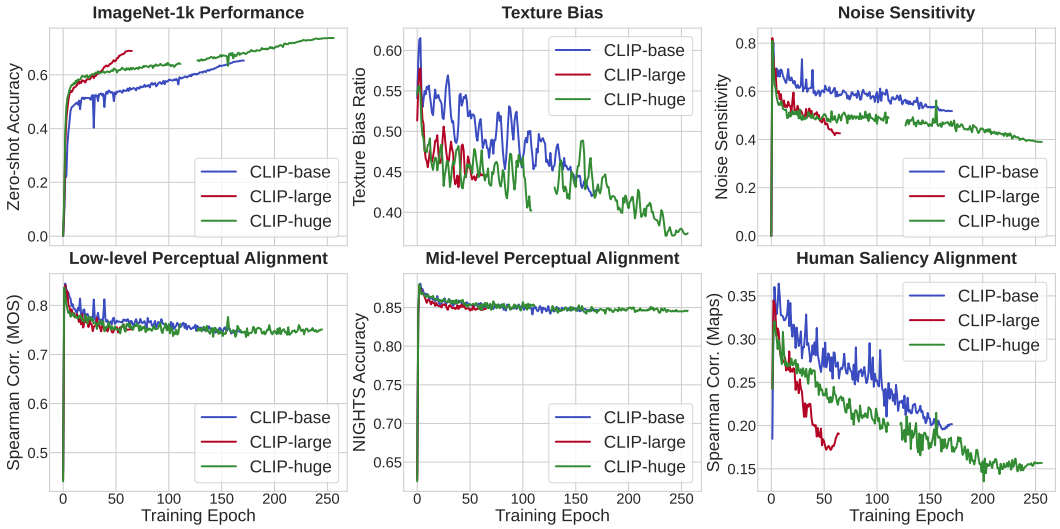

Figure 1: Evolution of CLIP representations across Base, Large, and Huge model sizes. **Top row:** Zero-shot ImageNet-1k accuracy, model texture bias, and noise sensitivity (normalized accuracy drop). **Bottom row:** Spearman correlation with Mean Opinion Scores (Low-level), NIGHTS accuracy (Mid-level), and saliency alignment (MIT1003).

This representational shift was accompanied by coordinated changes in noise robustness (Fig. 1, top-right) and perceptual alignment (Fig. 1, bottom row). The early training stages, characterized by strong texture bias, corresponded to a peak alignment with human judgments of low-level distortion (Mean Opinion Score, MOS) and with human visual attention (MIT1003). However, during this same period, models were most sensitive to Gaussian noise perturbations. As training progressed and representations become increasingly shape-oriented, noise robustness improved, while alignment with low-level perceptual metrics gradually declined.

Together, these results reveal a systematic co-evolution of texture bias, perceptual alignment, and noise robustness during CLIP training. The same qualitative dynamics were observed across all model sizes, indicating that this trade-off between early perceptual sensitivity and later robustness to noise is a general property of CLIP, rather than a scale-dependent effect.

3.2 Quantitative Correlation Analysis

The shared developmental trajectory of these properties is further supported by Spearman correlation analyses computed independently for all evaluated model sizes (ViT-Base-16, ViT-Large-14, and ViT-Huge-14). Across all scales, we observed highly consistent correlation patterns, indicating that the relationships between feature bias, noise robustness, classification accuracy, and human alignment metrics are largely invariant to model size. Table 1 reports the correlation matrix for the ViT-Base-16 model which is representative of the correlations found in the other models. These correlations were calculated by pairing the values of each metric recorded at every training epoch, from random initialization to the final available checkpoint. By using Spearman's rank correlation, we quantified the strength and direction of the monotonic relationships between variables throughout

the training process. The results of texture bias, noise sensitivity, and all measures of human alignment are strongly positively correlated ($r \geq 0.7$ for all pairs), and all are strongly anti-correlated with zero-shot classification accuracy. These results quantitatively confirm that improvements in noise robustness and semantic accuracy emerge alongside a systematic reduction in texture-based, perceptually aligned representations.

Table 1: Spearman correlations between the analyzed metrics during the training for CLIP-base16. Similar correlations are obtained for the other models.

	ImageNet-1k	**Texture-bias**	**Noise sens.**	**Low-level**	**Mid-level**	**Saliency**
ImageNet-1k	1	-0.79	-0.90	-0.89	-0.81	-0.92
Texture-bias		1	0.78	0.75	0.70	0.77
Noise sens.			1	0.89	0.79	0.88
Low-level				1	0.79	0.87
Mid-level					1	0.86

4 Discussion and Conclusions

Our results provide an explicit characterization of how CLIP visual representations evolve throughout training, revealing a structured but non-monotonic relationship between feature bias, noise robustness, and multiple notions of human alignment. By moving beyond static analyses of final checkpoints, we show that representational properties relevant to both performance and biological plausibility emerge, peak, and deteriorate at different stages of training.

A central observation is the progressive reduction of texture bias accompanied by increasing shape sensitivity as training advances. This transition is consistently observed across all model sizes and co-occurs with improvements in zero-shot classification accuracy and robustness to Gaussian noise perturbations. Importantly, this shift toward reduced texture bias should be interpreted as a movement toward, but not a convergence to, human visual processing. Static evaluations of final checkpoints, including prior work on CLIP and ViTs, have already shown that even fully trained models remain substantially more texture-biased than humans Geirhos et al. (2021). Our results refine this picture by showing that training drives the model in a more shape-oriented direction, yet the final representations remain far from human-like in absolute terms.

Crucially, the reduction in texture bias does not translate into uniformly increased human alignment across all evaluated metrics. While later training stages exhibit improved classification performance and increased shape bias, both commonly associated with more human-like object recognition, alignment with human judgments of low-level image distortions and visual saliency systematically decreases over training. This indicates that texture bias, although opposite to human shape dominance in object recognition, is positively associated with several forms of perceptual and saliency alignment. Moreover, a high correlation between the perception of low-level distortions and saliency is consistent with the similarity of mechanisms in classical models of subjective distortion Teo & Heeger (1994); Laparra et al. (2010) and saliency Itti & Koch (2000); Li (2002).

The joint evolution of all six metrics reveals a structured trade-off rather than independent trends. Texture bias, noise sensitivity, and all alignment measures are strongly correlated throughout training, and all are strongly anti-correlated with classification accuracy. As representations become more abstract and robust, they also become less aligned with human judgments at low, mid, and attentional levels, despite becoming more shape-biased and more accurate in classification. This demonstrates that increasing alignment along one axis (shape bias and semantic decisions) can coincide with decreasing alignment along others, highlighting that "human alignment" is inherently multi-dimensional and cannot be inferred from a single metric. Finally, the temporal profiles of all metrics suggest that training has not reached a stable endpoint in representational terms. Several curves continue to evolve steadily at the final available checkpoints, without clear saturation.

In conclusion, our findings demonstrate that representational evolution is a fundamental component of CLIP training. The model becomes progressively more shape-biased, more accurate, and more robust, yet remains far from human vision and simultaneously diverges from humans in several perceptual and attentional dimensions. These results underscore the importance of analyzing training dynamics rather than relying solely on final checkpoints, and suggest that future work should explicitly regulate representational trajectories if the goal is to achieve broader and more consistent alignment with human visual processing.

ACKNOWLEDGMENTS

This work was supported in part by MICIIN/FEDER/UE under Grant PID2023-152133NB-I00; in part by Spanish MIU under Grant FPU21/02256; in part by Generalitat Valenciana under Project GV/2021/074, CIPROM/2021/056; and in part by the grant BBVA Foundations of Science program: Maths, Stats, Comp. Sci. and AI (VIS4NN). The authors gratefully acknowledge the computer resources at Artemisa and the technical support provided by the European Union through the 2014-2020 ERDF Operative Programme of Comunitat Valenciana, project IDIFEDER/2018/048.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020.

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.

Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, James R. Glass, LIFEI HUANG, Jason E Weston, Luke Zettlemoyer, Xinlei Chen, Zhuang Liu, Saining Xie, Wen tau Yih, Shang-Wen Li, and Hu Xu. Meta CLIP 2: A worldwide scaling recipe. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=aYRNINhNGV.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Reza Esfandiarpoor, Cristina Menghini, and Stephen Bach. If clip could talk: Understanding vision-language model representations through their preferred concept descriptions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 9797–9819, 2024.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data, 2023.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*, 2018.

Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, 34:23885–23899, 2021.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

P. Hernández-Cámara, J. Vila-Tomás, J. Malo, and V. Laparra. Measuring human-CLIP alignment at different abstraction levels. In *ICLR 2024 Workshop on Representational Alignment*, 2024. URL https://openreview.net/forum?id=xQyhHjLGmj.

P. Hernández-Cámara, J. Vila-Tomás, V. Laparra, and J. Malo. Dissecting the effectiveness of deep features as metric of perceptual image quality. *Neural Networks*, 185:107189, 2025.

Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10):1489–1506, 2000. ISSN 0042-6989. doi: https://doi.org/10.1016/S0042-6989(99)00163-7.

Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *IEEE International Conference on Computer Vision (ICCV)*, 2009.

Hamid Kazemi, Atoosa Chegini, Jonas Geiping, Soheil Feizi, and Tom Goldstein. What do we learn from inverting clip models? In *NeurIPS 2024 Workshop on Safe Generative AI*, 2024. URL https://openreview.net/forum?id=GD23U1SWik.

M. Kumar, N. Houlsby, N. Kalchbrenner, and E. Cubuk. Do better imagenet classifiers assess perceptual similarity better? *Trans. Mach. Learn. Res.*, 2022, 2022.

Barbara Landau, Linda B Smith, and Susan S Jones. The importance of shape in early lexical learning. *Cognitive development*, 3(3):299–321, 1988.

V. Laparra, J. Muñoz-Marí, and J. Malo. Divisive normalization image quality metric revisited. *J. Opt. Soc. Am. A*, 27(4):852–864, Apr 2010. doi: 10.1364/JOSAA.27.000852.

Zhaoping Li. A saliency map in primary visual cortex. *Trends in Cognitive Sciences*, 6(1):9–16, 2002. ISSN 1364-6613. doi: https://doi.org/10.1016/S1364-6613(00)01817-9.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *2019 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–3. IEEE, 2019.

Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. Image database tid2013: Peculiarities, results and perspectives. *Signal processing: Image communication*, 30:57–77, 2015.

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*, pp. 8748–8763. PmLR, 2021.

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.

P.C. Teo and D.J. Heeger. Perceptual image distortion. In *Proc. IEEE Int. Conf. Im. Proc.*, volume 2, pp. 982–986 vol.2, 1994. doi: 10.1109/ICIP.1994.413502.

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*, pp. 11975–11986, 2023.