Evolution of Low-Level and Texture Human-CLIP Alignment

Pablo Hernández-Cámara[a*], **Jose Manuel Jaén-Lorites**[b], **Jorge Vila-Tomás**[a] **Jesus Malo**[a], **Valero Laparra**[a]
[a] Image Processing Lab, Universidad de Valencia, Paterna, Spain
[b] Center for Biomaterials and Tissue Engineering Universitat Politecnica de Valencia, Valencia, Spain
[*] Corresponding author: pablo.hernandez-camara@uv.es

Abstract

During the training of multi-modal models like CLIP, we observed an intriguing phenomenon: the correlation with low-level human image quality assessments peaks in the early epochs before gradually declining. This study investigates this observation and seeks to understand its causes through two key factors: shape-texture bias alignment and classification accuracy drop under noise. Our findings suggest that CLIP initially learn low-level visual features, enhancing its alignment with low-level human perception but also increasing its sensitivity to noise and its texture bias. As training progresses, the model shifts toward more abstract shape-based representations, improving noise robustness but reducing alignment with low-level human perception. These results suggest that these factors shared an underlying learning mechanism and provide new insights into optimizing the trade-off between perceptual alignment and robustness in vision-language models.

Keywords: Vision-Language Models; CLIP; Perceptual Dynamics; Texture-Shape Bias; Noise Robustness; Human Alignment

Introduction

Contrastive Language-Image Pretraining (CLIP) (Radford et al., 2021) models shown impressive zero-shot learning capabilities by aligning visual and linguistic representations. However, how these representations evolve during training remains unclear, particularly regarding low-level visual information, semantic features, and their alignment with human perception.

Human perception relies on a hierarchy of visual features, from low-level cues (e.g., texture and color) to higher-level semantic representations (DiCarlo, Zoccolan, & Rust, 2012). Previous work has shown that convolutional neural networks (CNNs) exhibit a strong bias toward texture, while vision transformers (ViTs) show a greater bias toward shape-based representations (Geirhos et al., 2018; Naseer et al., 2021). However, understanding how these biases evolve during training in multi-modal models like CLIP remains an open question.

Our study began by examining the model's alignment with low-level human perception, specifically its correlation with human image quality assessments. Surprisingly, we observed that this correlation does not steadily increase throughout training epochs but instead peaks early in training before declining. But why does this happen? This unexpected behavior led us to investigate its underlying causes. To address this question, we analyze two potential explanations: (1) shape-texture bias, which describes the model's preference for textures or shapes when classifying images, and (2) sensitivity to noise, measured through the relative accuracy drop when Gaussian perturbations are introduced into images.

Our results reveal that during the early training stages, CLIP aligns closely with human perception because it emphasizes low-level features such as textures and local patterns. However, this initial reliance on textures also makes the model more vulnerable to noise. As training progresses, its visual representations become more abstract and shape-based, improving robustness but reducing alignment with low-level human perception. This study provides a novel perspective on the learning dynamics of multi-modal models, highlighting the interplay between perceptual alignment, shape-texture bias, and noise robustness. Understanding this transition is crucial for improving future vision-language architectures, aiming for an optimal balance between human perceptual alignment and robustness in challenging conditions.

Methods

We analyze the OpenCLIP ViT-Base16 model throughout its training trajectory, using checkpoints from epoch 0 (random initialization) to epoch 65 (Cherti et al., 2023). This step-by-step analysis allows us to track how the model's alignment with human perception evolves over the epochs. Our study began by examining 2 factors: 1) Cifar100 0-shot classification accuracy and 2) low-level human perceptual alignment, measuring the correlation between CLIP's feature representations and human image quality assessments. Surprisingly, we found that while the classification accuracy always increases with the epochs, the low-level correlation peaks early in training before declining. To understand this phenomenon, we investigate two key factors: 3) shape-texture bias alignment and 4) sensitivity to noise. Therefore, we evaluate three key dimensions:

- **1) Classification accuracy:** We evaluate the zero-shot classification accuracy on Cifar100. For each image, we compute the similarity between the image representation and the representation of the texts *An image of* $\{$*class*$\}$.
- **2) Low-level Human Alignment:** We measure human perceptual alignment using the TID2013 image quality assessment database (Ponomarenko et al., 2015). For each epoch, we compute the similarity between each pair of original and distorted images in the model's embedding space. We then correlate these similarity scores with the human Mean Opinion Score (MOS) to quantify the alignment between model predictions and human perception.

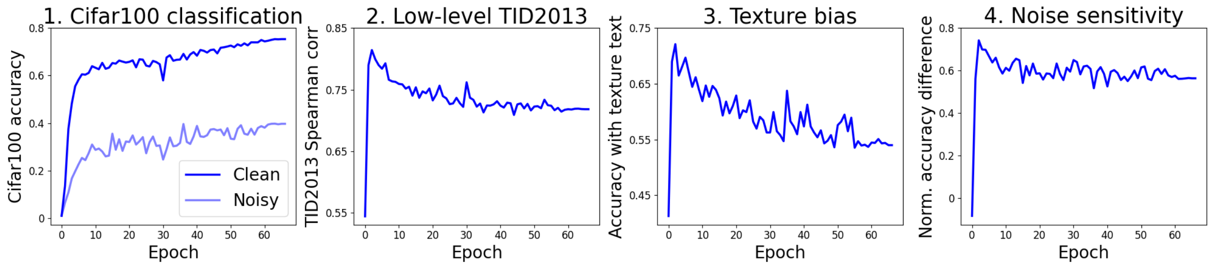

Figure 1: **CLIP alignment evolution.** First: Clean and noisy zero-shot Cifar100 accuracy. Second: low-level TID2013 correlation. Third: Model texture bias on texture-shape conflict images. Fourth: Cifar100 normalized accuracy drop due to Gaussian noise.

- **3) Texture-Shape Bias:** We quantify the model texture bias using the Geirhos Texture-Shape Bias dataset (Geirhos et al., 2018). For each conflict image (an image with a shape from one class and a texture from another), we compute the image's similarity with two textual descriptions corresponding to the shape and texture classes. We classify the image based on which text has higher similarity, determining whether the model favors shape or texture.
- **4) Noise sensitivity:** We evaluate the model's noise sensitivity by measuring the relative accuracy drop when Gaussian noise is introduced. Specifically, we compare the zero-shot classification accuracy on the clean CIFAR-100 images with the accuracy on the corrupted CIFAR-100-C dataset with Gaussian noise (Hendrycks & Dietterich, 2019). The relative accuracy drop reflects the model's sensitivity to image perturbations.

For each epoch, we compute these three metrics and zero-shot classification accuracy to track how CLIP's internal representations evolve throughout training. This analysis helps us uncover key transitions in the model's behavior, particularly the shift from early reliance on low-level, texture-based features to more abstract and robust representations. By examining these trends, we aim to explain why the alignment with human perception initially peaks before declining.

Results

Figure 1 illustrates how the different measured factors evolve during the epochs. In the first panel, we see that zero-shot Cifar100 classification accuracy always increases with epochs. As expected, when evaluated on the noisy images, the accuracy is lower but still increases with epochs. In the second panel, we examined low-level human alignment, finding that the correlation with human perceptual scores is highest in the initial epochs before gradually declining. This suggests that early in training, the model relies more on low-level visual features, but as it learns, it shifts toward more abstract representations. The third panel shows the evolution of texture-shape bias. The model exhibits a strong texture bias at the beginning of training, which progressively decreases over time. This indicates a transition from texture-based representations toward more shape-based abstractions. Finally, the last panel shows the robustness to noise analysis. It reveals that the model experiences the highest accuracy drop due to Gaussian noise in the early epochs, meaning it is initially more sensitive to perturbations. Over time, this sensitivity decreases, reflecting an improvement in robustness as the model's feature representations become more abstract.

	Acc.	TID	Texture
TID	-0.795		
Texture	-0.737	0.852	
Noise sens.	-0.441	0.836	0.618

Table 1: Pearson correlation between the four curves: accuracy on Cifar100, TID2013 correlation, texture bias and normalized accuracy drop due to Gaussian noise.

The three metrics follow a consistent pattern: a peak during the initial training stages, followed by a gradual decline. Table 1 show the Pearson correlations between these curves and further support this relationship.

These aligned dynamics suggest a shared underlying mechanism where the model starts by emphasizing low-level texture-based features and moving toward more shape-biased abstract and robust representations as training progresses.

Conclusions

Our analysis reveals a shared dynamic across low-level human alignment, texture bias, and noise robustness in CLIP's training. All three metrics peak in the early epochs before declining, suggesting an initial reliance on low-level, texture-based features that later shifts toward more abstract, shape-based representations and greater robustness. This highlights a trade-off between early perceptual alignment and and later robustness. The initial reliance on low-level features enhances texture bias and increases vulnerability to noise, while further training shifts the model toward more abstract, shape-based representations and greater robustness.

These findings provide new insights into CLIP's learning dynamics and raise questions about whether similar patterns occur in other multi-modal models. Future research could explore architectural adjustments or training strategies to balance early-stage perceptual alignment with later robustness. Extending this analysis to other vision-language models may also reveal whether this trend is a general phenomenon.

From a practical perspective, optimizing this transition could improve applications requiring fine-grained visual reasoning or robustness in noisy environments, guiding the development of models that better align with human perception while maintaining strong generalization.

Acknowledgments

This work was supported in part by MCIN/AEI/FEDER/UE under Grants PID2020-118071GB-I00 and PID2023-152133NB-I00, by Spanish MIU under Grant FPU21/02256 and in part by Generalitat Valenciana under Projects GV/2021/074, CIPROM/2021/056, and by the grant BBVA Foundations of Science program: Maths, Stats, Comp. Sci. and AI (VIS4NN). Some computer resources were provided by Artemisa, funded by the EU ERDF through the Instituto de Física Corpuscular, IFIC (CSIC-UV).

References

Cherti, M., et al. (2023). Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 2818–2829).

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434.

Geirhos, R., et al. (2018). Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations.*

Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*.

Naseer, M. M., et al. (2021). Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, *34*, 23296–23308.

Ponomarenko, N., et al. (2015). Image database tid2013: Peculiarities, results and perspectives. *Signal processing: Image communication*, *30*, 57–77.

Radford, A., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).