Interpreting style–content parsing in vision–language models

Fan L. Cheng
Columbia University
New York, NY 10027
fan.cheng@columbia.edu

Xin Jing
AstraBreeze, Inc.
New York, NY 10011
xin_jing@mfe.berkeley.edu

Abstract

Style refers to the distinctive manner of expressing content, and humans can both recognize content across stylistic transformations and detect stylistic consistencies across different contents. Prior work has shown that vision–language models (VLMs) exhibit steerable texture–shape biases, with language supervision shifting this tradeoff at the behavioral level. However, the internal representational dynamics of style and content—how they emerge across layers and how language pathways modulate them—remain poorly understood. Here, we adapt neuroscience-inspired tools to dissect style and content representations in a large VLM. We show that vision encoders strongly preserve stylistic signals while progressively enhancing content selectivity, and that language pathways further amplify content representations at the expense of style. Prompting can modestly steer these balances, but content remains dominant in deeper layers. These findings provide systematic evidence of style–content dissociation in multimodal models, guiding the design of architectures that more effectively balance style and content.

1 Introduction

Neural networks encode both style and content. Early evidence that artificial neural networks capture stylistic regularities came from style recognition using CNN features [1], soon followed by the breakthrough of neural style transfer: Pretrained CNNs were used to separate and recombine content and style by matching Gram-matrix statistics or covariances of feature maps [2–4]. A rich literature then optimized for speed, controllability, and generality [5–9]. StyleGAN explicitly modulates layers by a latent "style" to factor high-level attributes from stochastic detail [10]. Comparative surveys now document how CNN- and Transformer-based systems differ for style transfer and manipulation [11].

More recently, studies revealed that CNNs are more texture-biased than Vision Transformers (ViTs) [12, 13]. Rather than relying on convolutional locality, ViTs use global self-attention, yielding different internal progressions and inductive biases. Analyses show flatter specialization across layers and stronger propagation of low-level information [14], enhanced reliance on global shape with reduced texture bias and strong occlusion resilience [15], and distinct corruption-robustness patterns compared with CNNs and MLP-Mixers that reflect architectural inductive biases [13]. Training can tune these biases; e.g., shape–texture debiasing reduces single-cue reliance [16]. Frequency-domain studies argue multi-head attention tends toward low-pass, global structure, while convolutions emphasize high-frequency local detail [17]. Self-supervised ViTs exhibit emergent grouping and token semantics [18], and recent evidence suggests ViTs explicitly encode relations among objects—an aspect of "content" beyond shape alone [19].

VLMs are typically more shape-biased than their vision encoders and that this bias is steerable by language—e.g., prompting alone can move shape bias from ~49% to ~72% (humans ~96%)—with

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: CogInterp: Interpreting Cognition in Deep Learning Models.

complementary visual interventions (noise, patch shuffling) [20]. Contamination-controlled splits further reveal that CLIP's strong results on stylized/sketch benchmarks partly reflect exposure to rendition images during pretraining rather than true domain generalization[21].

These findings highlight the influence of language on style–content balances, but they are based on model behavior rather than internal mechanisms. It remains unclear where style information is preserved or discarded inside VLMs, how separable style and content are across network depth, and how multimodal supervision reshapes these representations. We address these questions by directly probing internal geometry and decodability of layerwise representations using a controlled set of styled images, thereby bridging psychophysics of style perception and neural network interpretability.

2 Methods

2.1 Dataset

We use a controlled image set in which **content** (e.g., beaches, libraries, mountains, bedrooms) and **style** (e.g., Van Gogh, Monet, Klimt, Munch, Pollock) vary orthogonally (Figure 1). Style variations were generated by applying neural style transfer to naturalistic scene photographs (images available on OSF: `https://osf.io/mb3nh/`; [22]).

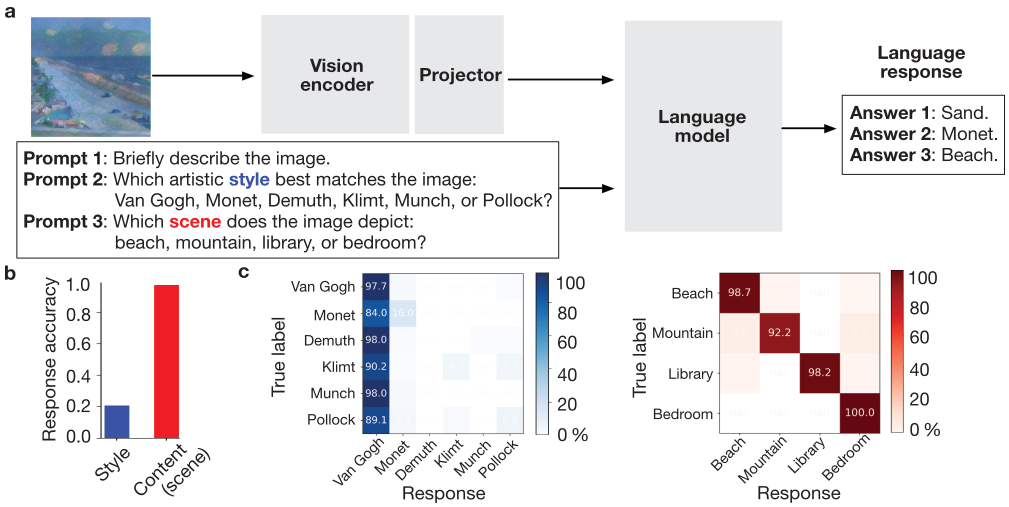

Figure 1: Prompting framework and behavioral evaluation of a vision–language model (VLM). (a) Three prompt types were used: free-form description (prompt 1), closed-set style classification (prompt 2), and closed-set content classification (prompt 3). (b) Overall response accuracy for style and content. (c) Confusion matrices for style (blue) and content (red) categories.

2.2 Model and prompting framework

We use the pretrained LLaVA model Liu et al. [23], which connects a pretrained CLIP ViT-L/14 vision encoder [24] to a large language model via a learned projection layer, enabling end-to-end multimodal instruction following (Figure 1a). Training proceeds in two stages: first, the projection layer is aligned by matching image features to paired captions; second, the entire model is fine-tuned with multimodal instruction data to enable conversational use. Following this two-stage approach, we assume that all modules are fine-tuned for multimodal interaction.

We use the same prompt set for all analyses (behavioral, RSA, and probes), so that the same linguistic framing also defines the language-conditioned activations we analyze. For closed-set prompts, we constrain decoding to the candidate label set (exact match on allowed options; minor variants are normalized) and compute accuracy. The same prompts are used when extracting language features while the model processes the image–prompt pair. These prompt-conditioned activations provide the basis for the representational analyses described below, enabling us to examine how style and content information evolve across layers under different forms of linguistic guidance.

2.3 Representational analysis

Representational Similarity Analysis (RSA). We characterize layerwise representational geometry using RSA [25, 26]. For each image–prompt condition, let $h_{\ell i} \in \mathbb{R}^D$ denote the activation vector at layer ℓ for stimulus i ($i = 1, \ldots, N$). After mean-centering each feature across stimuli, we compute a representational dissimilarity matrix (RDM) $R_\ell \in \mathbb{R}^{N \times N}$ with *correlation distance*.

$$R_\ell(i, j) \;=\; 1 - \mathrm{corr}\big(h_{\ell i},\, h_{\ell j}\big),$$

We define binary target RDMs for *style* and *content* using the Kronecker delta $\delta(\cdot, \cdot)$:

$$T^{\text{style}}_{ij} = 1 - \delta(s_i, s_j), \qquad T^{\text{content}}_{ij} = 1 - \delta(c_i, c_j),$$

where s_i and c_i denote the style and content labels of image i, respectively. Thus, within-class pairs receive 0 and between-class pairs receive 1 (diagonals are 0 by definition). We quantify alignment via rank correlation (Spearman) between layer-wise and target geometries. This procedure reveals whether layers preferentially organize images by style or by content, and how this organization evolves with depth and linguistic guidance.

Linear Probes. To quantify the accessibility of style and content information at each layer [27], we extract activations in response to the image–prompt pairs. We first apply principal component analysis (PCA) to reduce dimensionality (explaining 95% variance), then split the dataset into training (80%) and test (20%) partitions and train an ℓ_2-regularized logistic regression classifier. Probes are trained separately for *content* classification and for *style* classification. Probe accuracy provides a measure of how much linearly decodable information about content or style is present at each layer.

3 Results

3.1 Behavioral responses under style and content prompts

We first assessed the VLM's outputs directly by presenting images under three prompts (free-form description, closed-set style classification, closed-set content classification; Figure 1a). The model produced descriptions that consistently identified (part of) the content, but showed marked differences in accuracy across style and content tasks. Response accuracy was near ceiling for content, but remained close to chance for style (Figure 1b). The confusion matrices further illustrate this asymmetry (Figure 1c). For content prompts, responses aligned closely with ground-truth labels. In

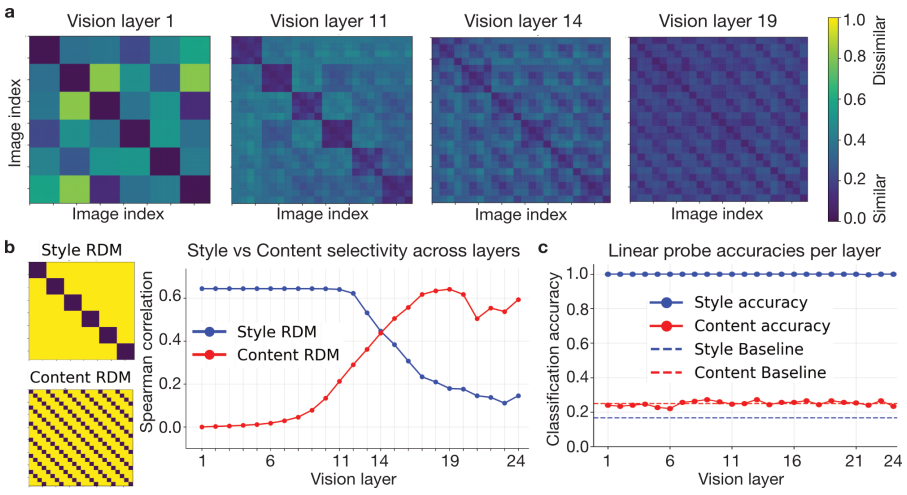

Figure 2: **Layerwise representational analyses of the vision encoder.** (a) Example representational dissimilarity matrices (RDMs). (b) RSA correlations between layer-wise RDMs and target RDMs for style and content. (c) Linear probe accuracy.

contrast, style prompts elicited systematic confusions, with the model defaulting disproportionately to Van Gogh, while misclassifying other styles.

3.2 Style and content selectivity in vision layers

We next examined the internal geometry of the vision encoder, which is independent of prompts. Early layers yield block-diagonal patterns consistent with stylistic grouping, whereas later layers become increasingly homogeneous (Figure 2a). Across the first ten layers, network RDMs correlated strongly with the style target matrix (Figure 2b), indicating that low- and mid-level layers cluster images by style. Beyond layer 12, style correlations declined as content correlations rose, with late layers primarily organizing by scene category. Style labels were linearly decodable from every layer, achieving near-perfect accuracy even early in the encoder (Figure 2c). By contrast, content classification lagged near baseline across layers.

Reconciling RSA and probe results. RSA and linear probe analyses appear to diverge: RSA shows that later vision layers have similar representations for images with the same content but different styles, whereas probes reveal that style remains linearly decodable throughout the network and content decoding remains comparatively weak. These results can be reconciled by noting that RSA captures the *dominant geometry* of the representational space—whether inter-image distances align with style or content—whereas probes test the *presence of linearly accessible information*. In later layers, the geometry is organized primarily by content, so style variation no longer drives representational distances. However, residual style features persist in the embedding vectors, allowing probes to classify them with high accuracy even when they no longer dominate similarity structure. Conversely, although content governs the overall geometry, it may be encoded in a distributed or nonlinear fashion that complicates linear decoding. Thus RSA and probes jointly reveal that style is consistently available but deemphasized in later layers, while content increasingly shapes representational geometry without becoming trivially linearly separable.

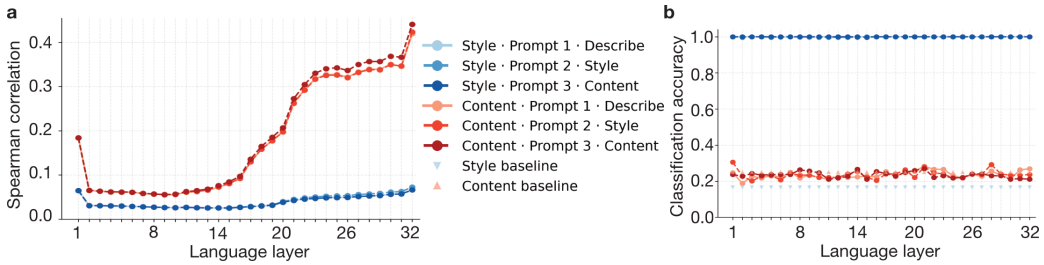

Figure 3: **Layerwise representational analyses of the language layers.** (a) RSA correlations between layer-wise and target RDMs. (b) Linear probe accuracy.

3.3 Prompting effects on style–content dissociation in language layers

RSA correlations at the first language layer were much lower than those observed in the final vision layer (Fig. 3a). Content correlations increased markedly with layer depth, showing a sharp rise beginning around layer 18. By contrast, style correlations remained low, with only modest increases in later layers. These trends were robust across prompt formulations. Moreover, content- and style-oriented prompts slightly steered the representational geometry, biasing it toward preferential organization by content or style, respectively. Linear probes at the last layer do *not* reach the high content accuracy observed in behavioral outputs (Fig. 3b), likely because the final hidden layer was optimized for next-token prediction rather than for linearly separating style and content categories.

4 Discussion

Our analyses reveal a systematic dissociation between style and content representations in vision–language models. Vision encoders robustly preserve stylistic information alongside content, but language pathways progressively reorganize the representational space to emphasize content semantics at the expense of style. Whereas prompting can steer model behavior toward style or content, representational analyses show that the underlying geometry is already strongly biased toward content in later layers. Our results suggest that current architectures may undervalue stylistic information. More balanced VLMs may require mechanisms that preserve stylistic signals alongside semantic abstraction, enabling models to flexibly engage with both what is depicted and how it is expressed.

Acknowledgments and Disclosure of Funding

The authors declare no competing interests.

References

[1] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. Recognizing image style. *arXiv preprint arXiv:1311.3715*, 2013.

[2] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

[3] Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. *Advances in neural information processing systems*, 28, 2015.

[4] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

[5] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[6] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. *arXiv preprint arXiv:1701.01036*, 2017.

[7] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.

[8] Golnaz Ghiasi, Honglak Lee, Manjunath Kudlur, Vincent Dumoulin, and Jonathon Shlens. Exploring the structure of a real-time, arbitrary neural artistic stylization network. *arXiv preprint arXiv:1705.06830*, 2017.

[9] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018.

[10] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[11] Hua-Peng Wei, Ying-Ying Deng, Fan Tang, Xing-Jia Pan, and Wei-Ming Dong. A comparative study of cnn-and transformer-based visual style transfer. *Journal of Computer Science and Technology*, 37(3):601–614, 2022.

[12] 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.

[13] Katelyn Morrison, Benjamin Gilby, Colton Lipchak, Adam Mattioli, and Adriana Kovashka. Exploring corruption robustness: Inductive biases in vision transformers and mlp-mixers. *arXiv preprint arXiv:2106.13122*, 2021.

[14] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128, 2021.

[15] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34:23296–23308, 2021.

[16] Yingwei Li, Qihang Yu, Mingxing Tan, Jieru Mei, Peng Tang, Wei Shen, Alan Yuille, and Cihang Xie. Shape-texture debiased neural network training. *arXiv preprint arXiv:2010.05981*, 2020. doi: 10.48550/arXiv.2010.05981. URL `https://arxiv.org/abs/2010.05981`. ICLR 2021.

[17] Namuk Park and Songkuk Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022.

[18] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[19] Michael Lepori, Alexa Tartaglini, Wai Keen Vong, Thomas Serre, Brenden M Lake, and Ellie Pavlick. Beyond the doors of perception: Vision transformers represent relations between objects. *Advances in Neural Information Processing Systems*, 37:131503–131544, 2024.

[20] Paul Gavrikov, Jovita Lukasik, Steffen Jung, Robert Geirhos, Bianca Lamm, Muhammad Jehanzeb Mirza, Margret Keuper, and Janis Keuper. Are vision language models texture or shape biased and can we steer them? *arXiv preprint arXiv:2403.09193*, 2024.

[21] Prasanna Mayilvahanan, Roland S Zimmermann, Thaddäus Wiedemer, Evgenia Rusak, Attila Juhos, Matthias Bethge, and Wieland Brendel. In search of forgotten domain generalization. *arXiv preprint arXiv:2410.08258*, 2024.

[22] Tal Boger and Chaz Firestone. The psychophysics of style. *Nature Human Behaviour*, pages 1–13, 2025.

[23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. URL `https://arxiv.org/abs/2304.08485`.

[24] 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.

[25] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

[26] Jörn Diedrichsen and Nikolaus Kriegeskorte. Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS computational biology*, 13(4):e1005508, 2017.

[27] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.