Convergence of Spectral Principal Paths: How Deep Networks Distill Linear Representations from Noisy Inputs

Anonymous Author(s)
Affiliation
Address
email

Abstract

High-level representations have become a central focus in enhancing AI transparency and control, shifting attention from individual neurons or circuits to structured semantic directions that align with human-interpretable concepts. Motivated by the Linear Representation Hypothesis (LRH), we propose the Input-Space Linearity Hypothesis (ISLH), which posits that concept-aligned directions originate in the input space and are selectively amplified with increasing depth. We then introduce the Spectral Principal Path (SPP) framework, which formalizes how deep networks progressively distill linear representations along a small set of dominant spectral directions. Building on this framework, we further demonstrate the multimodal robustness of these representations in Vision-Language Models (VLMs). By bridging theoretical insights with empirical validation, this work advances a structured theory of representation formation in deep networks, paving the way for improving AI robustness, fairness, and transparency.

1 Introduction

Deep learning has achieved remarkable success across various domains, including computer vision [1], natural language processing [2], and speech recognition [3, 4]. However, the internal mechanisms of neural networks remain opaque. Despite advances in visualization and interpretability techniques, the transformation of inputs into high-level representations and the interactions among neurons are still not fully understood [5, 6, 7]. This lack of transparency leads to the characterization of neural networks as "black boxes" [5, 6], raising concerns about their reliability, particularly in high-stakes applications such as healthcare [8], finance [9], and legal decision-making [10].

Previous works have demonstrated the potential of representations as a new perspective on AI transparency. For example, neural networks trained to play chess exhibit internal representations of board positions and strategies [11]. Similarly, both generative and self-supervised models have been shown to develop emergent representations, such as semantic segmentation in vision tasks [12, 13]. Zou et al. [14] further formalized Representation Engineering (RepE), emphasizing its ability to extract meaningful concepts from a model's internal structure and control model behavior. RepE has emerged as a top-down approach to enhance the model transparency that focuses on representations rather than individual neurons or circuits, providing a more structured understanding of AI transparency and control. Another important contribution is the Linear Representation Hypothesis (LRH) [15]: as depth increases, task-relevant concepts become nearly linearly separable in the model's latent space, making them accessible with simple probes or linear edits.

Despite these promising advances, existing works on representations remain largely observational, relying on observed phenomena or intuitions. RepE identifies concept directions through contrastive

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

pairs (e.g., honesty vs. dishonesty), but explanations for why such directions emerge or remain stable across layers remain unexplored. Similarly, LRH assumes linearity in embedding and unembedding spaces, yet offers limited insight into why representations become linearly organized. These approaches typically do not address how representations scale or propagate through deep networks, leaving a gap in our understanding of their robustness, generality, and theoretical foundations.

In this work, we move beyond linear observations by introducing Spectral Principal Path (SPP) that explains the emergence and stability of linear representations in deep networks. We show that representations propagate through a small number of spectral principal paths—directions aligned with large singular values at each layer. This structure naturally explains why concept directions remain stable and linearly accessible across layers, offering a theoretical foundation for both RepE and the Linear Representation Hypothesis. We further extend this analysis to Vision-Language Models (VLMs), demonstrating how spectral dynamics govern the interaction between visual and linguistic modalities. Our framework not only bridges theory and practice but also provides concrete tools to improve robustness and interpretability in multimodal AI systems.

Our main contributions are as follows:

- **Input-Space Linearity Hypothesis (ISLH).** We extend the Linear Representation Hypothesis beyond embedding and unembedding spaces to the input space itself, showing that concept directions can be traced backward to the input space.
- **Spectral Principal Path (SPP)**. We propose a principled mechanism explaining how representations propagate and stabilize across layers via a small number of spectral principal paths—directions aligned with large singular vectors.
- **Multimodal robustness of representations**. We evaluate Representation Engineering in VLMs and demonstrate that linearly organized concept representations remain robust across modalities. This provides the first empirical validation of RepE's scalability in multimodal systems and supports the generality of spectral structure as a foundation for transparency and control.

2 Related Works

2.1 Representations in Neural Networks

Early work on word embeddings shows that neural networks can learn distributed representations that encode semantic relationships and compositional structures [16]. Follow-up studies [17, 18] further reveal that learned embeddings can implicitly encode abstract dimensions such as commonsense morality, even without explicit supervision. For instance, Radford et al. [18] observe that training a language model on product reviews results in the emergence of a sentiment-tracking neuron.

This phenomenon is not unique to language models. McGrath et al. [11] show that similar internal representations can be found in networks trained to play chess. In computer vision, recent studies [12, 13] demonstrate that both generative and self-supervised training objectives give rise to emergent semantic representations, such as those useful for segmentation tasks.

Building on this, Zou et al. [14] propose techniques to read and control these internal structures, including Linear Artificial Tomography (LAT) for extracting concept-aligned representations and methods for steering model behavior. Their study shows that RepE-style approaches can be used not only to detect but also to manipulate emergent properties, motivating more systematic efforts to characterize and intervene in high-level model behaviors. Theoretically, Park et al. [15] propose the Linear Representation Hypothesis: task-relevant concepts become nearly linearly separable in the model's latent space.

2.2 Approaches to Interpretability

Traditional interpretability techniques have focused on methods like saliency maps [19, 20, 21, 22], feature visualization [23, 21] and mechanistic interpretability [24, 25, 26]. Saliency maps [19] highlight important input regions by tracking gradients or activation values, yet they are often unstable and provide limited insight into the distributed nature of representations. Similarly, feature visualizations [23, 21] optimize inputs to activate specific neurons, but they may overlook the global

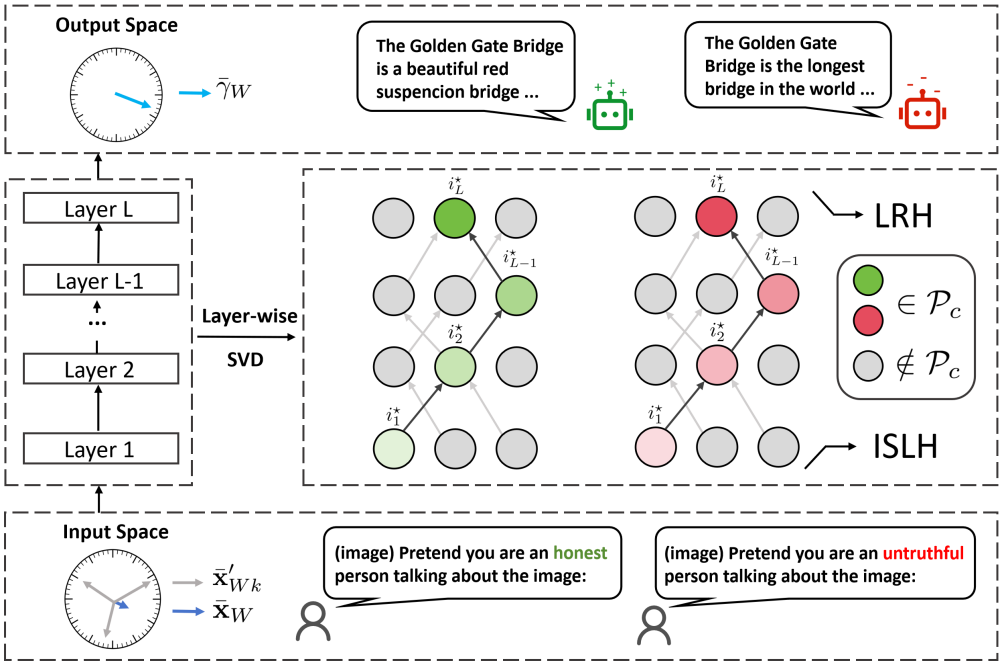

Figure 1: The overview of Spectral Principal Path framework.

structure of the emergent representations. Mechanistic interpretability [14] seeks to fully reverse engineer neural networks into their "source code", but the considerable manual effort and the difficulty of theoretically explaining neural networks as discrete circuits hinder their explainability.

In contrast, recent advances in interpretability have shifted the focus toward analyzing representation spaces. This top-down approach seeks to uncover high-level semantic directions that correspond to complex phenomena such as honesty, fairness, or bias. By extracting and analyzing these internal representations, researchers have opened new avenues to understand how large-scale AI models encode and preserve crucial information across layers, leading to more robust and interpretable AI systems.

3 Preliminaries

Linear Representation Hypothesis [15] We consider a concept W that has a *linear representation* in a model if there exists a vector $\bar{\gamma}_W$ in the unembedding space Γ and/or a vector $\bar{\lambda}_W$ in the embedding space Λ such that for any counterfactual pair $(Y(W=0), Y(W=1))$,

$$\gamma(Y(W=1)) - \gamma(Y(W=0)) \in \text{Cone}(\bar{\gamma}_W), \tag{1}$$

and for any context pair (λ_0, λ_1) that changes only W and not other causally separable concepts,

$$\lambda_1 - \lambda_0 \in \text{Cone}(\bar{\lambda}_W). \tag{2}$$

where $\text{Cone}(\mathbf{v}) = \{\alpha\mathbf{v} : \alpha > 0\}$, the embedding space is where input contexts are mapped to high-dimensional vectors before processing, capturing the model's internal representation of the input. The unembedding space is where each output token is represented, and predictions are made by computing inner products between input embeddings and output unembedding vectors. Unless stated otherwise, all discussions pertain to the embedding space Λ, as our goal is to trace how input linearity propagates through the network.

4 Spectral Principal Path Framework

The overview of the Spectral Principal Path framework is shown in Fig. 1. In the input space, the concept direction $\bar{\mathbf{x}}_W$ separates inputs with contrast concepts such as "honest" and "untruthful". As

activations propagate through the network, layer-wise SVD identifies spectral components, forming spectral principal paths $\mathcal{P}_c$. These dominant paths progressively amplify concept-relevant signals, leading to output representations linearly aligned with $\bar{\gamma}_W$.

4.1 Input-Space Linearity Hypothesis

Inspired by LRH, which uncovers linear concept axes in embedding and unembedding spaces, we take one step further and ask whether such axes already reside in the *raw input space*. Input-Space Linearity Hypothesis assumes that, in the *raw input space* $x \in \Upsilon$, there exists a discriminative direction $\bar{\mathbf{x}}_W$ such that

$$\mathbb{E}[x \mid W=1] - \mathbb{E}[x \mid W=0] \in \mathrm{Cone}(\bar{\mathbf{x}}_W), \tag{3}$$

yet each sample is an entangled mixture

$$x^{(i)} = \alpha_i \bar{\mathbf{x}}_W + \sum_{k=1}^{r} \beta_{i,k} \bar{\mathbf{x}}'_{Wk} + \varepsilon_i, \tag{4}$$

where $\{\bar{\mathbf{x}}'_{Wk}\}$ are spurious directions and ε_i is residual noise. ISLH states that for any intervention flipping only W, the induced input difference satisfies $\mathrm{Cone}(\bar{\mathbf{x}}_W)$.

ISLH pinpoints the origin of linearity by showing that concept axes already reside in raw input coordinates and are merely recovered and amplified during training; where training can be viewed as a noise-suppression process, where spectral principal paths with large singular values progressively dampen spurious components $\beta_{i,k}\bar{\mathbf{x}}'_{Wk}$; and, by grounding linearity at the input level, it becomes inherently modality-agnostic, extending Representation Engineering to multimodal models whose raw signals already encode task-relevant contrasts. Next, we will dive into the connection between ISLH and LRH:

Theorem 1 (ISLH sufficiency)**.** If the network satisfies the *Input-Space Linearity Hypothesis* (ISLH), and the representation dominates the cumulative gain $G(\mathcal{P})$ (shown in (12)), then its deep representations satisfy the *Linear Representation Hypothesis* (LRH); that is, concept classes become linearly separable in the latent space.

The proof is given in Appendix A.

4.2 Spectral Principal Path

We are now asking how such concept directions propagate through the network. While ISLH posits that concept-aligned directions already exist in the raw input space, it does not yet explain why these directions persist and become more prominent across layers. To address this, we introduce the *Spectral Principal Path* (SPP) framework, which shows that representations are distilled through a small set of principal spectral paths aligned with large singular vectors at each layer. This framework formalizes how ISLH leads to the emergence of the Linear Representation Hypothesis (LRH), providing a unified and mechanistically grounded view of representation stability.

Specifically, consider a generalized network

$$f_L(x) = W_L W_{L-1} \cdots W_1 x \;\equiv\; Mx, \qquad W_l \in \mathbb{R}^{d_l \times d_{l-1}}, \tag{5}$$

according to LRH, there exists a representation direction $\bar{\lambda}_W$, where the neural activity $f(x)$ can be linearly projected into that direction, formulating a representation score:

$$s(x) = \langle \bar{\lambda}_W,\ f_L(x) \rangle = \bar{\lambda}_W^\top M x, \qquad \bar{\lambda}_W \in \mathbb{R}^{d_L}. \tag{6}$$

While our theoretical formulation assumes a purely stacked linear architecture, we show our extension to residual connections and attention mechanisms. We provide a detailed discussion of these extensions in Appendix B.1.

Next we will calculate the back-propagated gradient of s using the chain rule,

$$\nabla_x s \;=\; \left(\prod_{l=1}^{L} \nabla f_{l \to (l-1)} \right)^\top \bar{\lambda}_W, \tag{7}$$

$$\nabla f_{l\to(l-1)} \;=\; W_l \;+\; \sum_k f_{l-1,k}\,\frac{\partial W_l}{\partial f_{l-1,k}}. \tag{8}$$

To make the structure of the layer-wise Jacobian in (8) explicit, we regard the gradient $\nabla f_{l\to(l-1)}$ as a matrix and apply its compact singular-value decomposition (SVD); this yields

$$\nabla f_{l\to(l-1)} = U^{(l)}\Sigma^{(l)}V^{(l)\top}, \quad \Sigma^{(l)} = \operatorname{diag}\big(\sigma_1^{(l)},\dots,\sigma_{r_l}^{(l)}\big), \tag{9}$$

therefore

$$\nabla_x s \;=\; V^{(1)}\Sigma^{(1)}U^{(1)\top}\cdots V^{(L)}\Sigma^{(L)}U^{(L)\top}\bar{\lambda}_W. \tag{10}$$

Unfolding the matrix products yields

$$\nabla_x s \;=\; \sum_{i_1,\dots,i_L}\Big(\prod_{l=1}^{L}\sigma_{i_l}^{(l)}\Big)V_{\cdot i_1}^{(1)}\Big(\prod_{l=1}^{L-1}\langle u_{i_l}^{(l)},\, V_{\cdot i_{l+1}}^{(l+1)}\rangle\Big)\langle u_{i_L}^{(L)},\, \bar{\lambda}_W\rangle\,, \tag{11}$$

where $\sigma_{i_l}^{(l)}$ is the signular value within the $\Sigma^{(l)}$ matrix, $u_{i_l}^{(l)}$ (resp. $V_{\cdot i_l}^{(l)}$) is the i_l-th left (resp. right) singular vector, and $\cdot i_l$ here means select the i_l-th column. Therefore (11) is dominated by paths whose cumulative gain is largest. Formally, we define Spectral Principal Path as follows:

Definition 1 (Spectral Principal Path)**.** Given the Jacobian decomposition across L layers $\nabla_x s$, each spectral path $\mathcal{P} = (i_1,\dots,i_L)$ contributes a cumulative gain given by

$$G(\mathcal{P}) := \Big(\prod_{l=1}^{L}\sigma_{i_l}^{(l)}\Big)V_{\cdot i_1}^{(1)}\Big(\prod_{l=1}^{L-1}\langle u_{i_l}^{(l)},\, V_{\cdot i_{l+1}}^{(l+1)}\rangle\Big)\langle u_{i_L}^{(L)},\, \bar{\lambda}_W\rangle\,, \tag{12}$$

the **Spectral Principal Path** is defined as $\mathcal{P}_c = (i_1^\star,\dots,i_L^\star)$ that maximizes the cumulative gain:

$$\mathcal{P}_c = (i_1^\star,\dots,i_L^\star) := \arg\max_{(i_1,\dots,i_L)} G(\mathcal{P}). \tag{13}$$

4.3 Connection between ISLH and SPP

To clarify how information specified by ISLH is propagated along an SPP, we introduce the notion of *spectral similarity* for a given spectral path $(i_1,\dots,i_L)$:

Definition 2 (Spectral Similarity)**.** For two consecutive layers l and $l{+}1$ in the unfolded Jacobian, and for indices (i_l, i_{l+1}), we define the *spectral similarity at layer* l as

$$\Theta(i_l, i_{l+1}) \;:=\; \big\langle u_{i_l}^{(l)},\, V_{\cdot i_{l+1}}^{(l+1)}\big\rangle, \qquad l = 1,\dots,L-1. \tag{14}$$

This quantity measures how well the i_l-th spectral component of layer l aligns with the i_{l+1}-th spectral component that enters layer $l{+}1$.

Empirically, we observe two coupled effects as the network reaches deeper.

1. **Stabilization of singular vectors.** As demonstrated in Fig. 2, the principal singular vectors $u_{i_*}^{(l)}$ change only marginally with f_l. Consequently, the spectral similarity of a few paths approaches 1, the stability of singular vectors implies that spectral similarity remains very high in deeper layers, allowing information to propagate consistently along those paths with high spectral similarity.
2. **Selective growth of singular values.** As demonstrated in Fig. 3, we observe that the singular values are growing as the layers deepen; and at the same depths only a very small subset of singular values $\sigma_{i_*}^{(l)}$ are amplified; the remainder stay close to their initial scale.

Putting these observations together shows that, in deep layers, the directions that (i) possess *large spectral similarity* and (ii) carry *large singular values* coincide, which satisfies dominant $G(\mathcal{P})$ condition. In other words, the network progressively funnels representation power into precisely those spectral directions that stay *globally aligned* across layers.

According to Theorem 1, this behaviour is exactly what the Input-Space Linearity Hypothesis (ISLH) predicts: concept-carrying directions are expected to form a low-dimensional subspace that is both *spectrally dominant* (large $\sigma_{i_*}^{(l)}$) and *structurally coherent* (large Θ) throughout the hierarchy, leading to LRH. Hence, the emergence of a handful of high-σ, high-similarity principal paths in SPP provides concrete spectral evidence in favour of the ISLH assumption.

5 Experiments

5.1 Experiment Setup

Dataset: We conduct our experiments on the Microsoft COCO (Common Objects in Context) dataset [27], a large-scale benchmark for vision-language tasks. COCO contains over 330K images, each with five human-annotated captions, covering diverse real-world scenes.

VLM: We employ Idefics2-8B [28, 29], a state-of-the-art VLM that extends the LLaMA architecture with a vision encoder, enabling multimodal reasoning over images and text. Idefics2-8B is designed for instruction-following, multimodal dialogue, and grounded language generation, making it an ideal candidate for studying conceptual representations in VLMs.

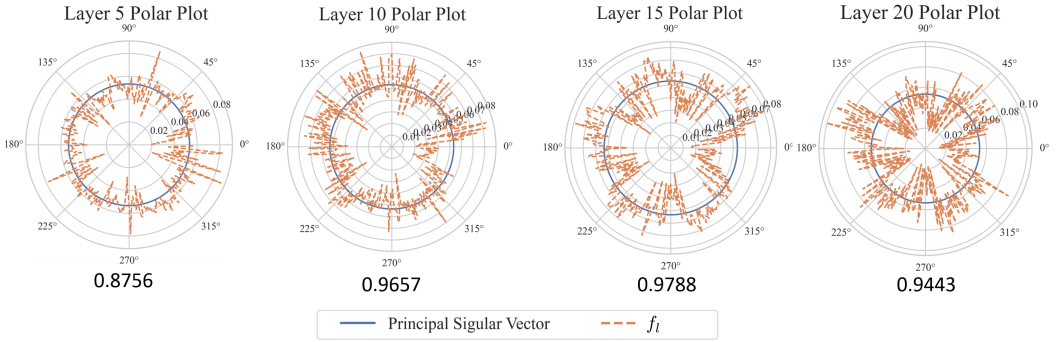

Figure 2: The polar plot demonstrates normalized connections between principal singular vector and f_l, where the number indicates their cosine similarity. The results showcase that f_l, especially in later layers, is very similar to the principal singular vector of that layer.

5.2 The Significant Alignment between Principal Singular Vector and f_l

Fig. 2 visualizes the connections between the principal singular vector, i.e., the singular vector with the largest singular value, and f_l. The results reveal a strong alignment between the principal singular vector and f_l, with their cosine similarity over 0.875. This experimental validation supports our theoretical claim that singular vectors with large singular values remain stable across layers, reinforcing their stability of spectral similarity.

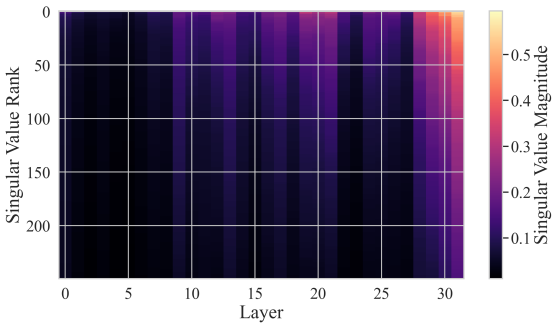

Figure 3: The singular value rank across layers.

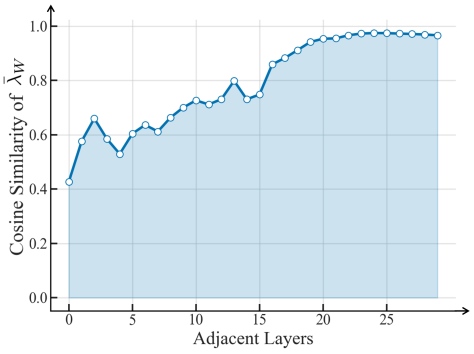

Figure 4: The cosine similarity of $\bar{\lambda}_W$ between adjacent layers.

5.3 Spectral Energy Concentration Across Layers

To investigate how spectral energy propagates through the network, we analyze the singular value spectrum of the layer-wise Jacobians. Fig. 3 presents a heatmap of the singular value magnitudes

across all layers. The x-axis indicates the layer index, and the y-axis corresponds to the ordered singular value indices. Color intensity reflects the magnitude of each singular value.

We observe that the singular values are growing as the layers deepen, and at the same depths, only a very small subset of singular values are amplified; the remainder stay close to their initial scale.

These results indicate that, with increasing depth, spectral energy becomes increasingly concentrated in a few dominant directions. Combined with the theoretical formulation in Section 4.3, this supports the hypothesis that high-magnitude spectral components dominate SPPs.

5.4 Inter-Layer Spectral Similarity of $\bar{\lambda}_W$

We further analyze the alignment of concept-carrying directions across adjacent layers by computing the average cosine similarity between the projections of $\bar{\lambda}_W$ at different layers. This measures how stable the representation direction remains as it propagates backward through the network. The results are shown in Fig. 4.

The curve demonstrates a clear upward trend: the inter-layer similarity of $\bar{\lambda}_W$ increases consistently with network depth, eventually approaching a value near 0.95 in the final layers. This suggests that the concept direction stabilizes as it propagates through deeper layers, aligning with the intuition of structured and coherent representation flow.

5.5 Multimodal Robustness of Representation

In this experiment, we explore the multimodal robustness of representation. Specifically, we analyze how VLMs encode fairness and honesty, and how these concepts persist or transform as information propagates through the model. These findings deepen our understanding of how representations enhance both interpretability and conceptual alignment in the context of multimodal reasoning.

5.5.1 Evaluating Honesty and Fairness in VLMs

To evaluate how well VLMs represent abstract ethical concepts, we analyze their handling of honesty and fairness in multimodal response generation. These concepts are critical for reducing misinformation and bias and serve as strong test cases for examining interpretability and ethical alignment in large-scale models.

To quantify this process, we compute token-wise honesty and fairness scores following RepE [14], measuring how closely activations align with concept directions at each layer. These results highlight the structured nature of ethical concept encoding in VLMs and support our broader claims about representation flow along spectral directions and its traceability from input to output.

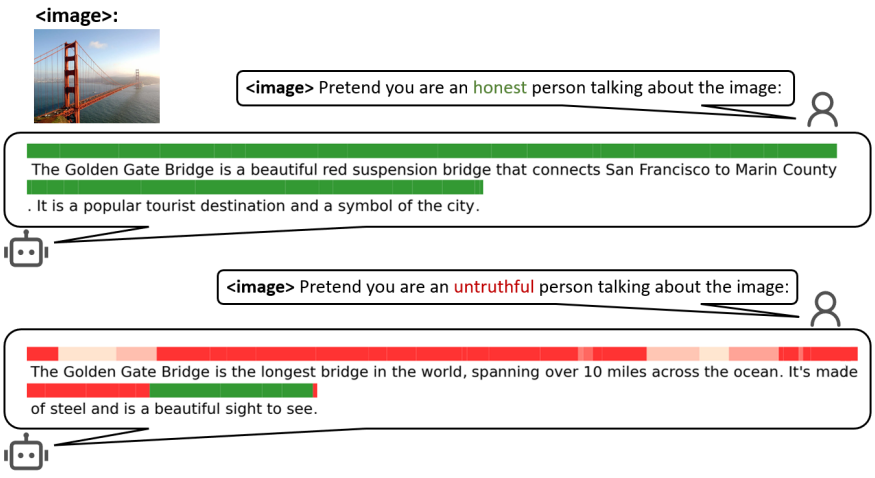

Figure 5: VLM response to an image of the Golden Gate Bridge with a prompt related to the concept of honesty, along with token-wise honesty scores. Green indicates high honesty, while red represents low honesty.

- **Honesty**: We define honesty as the model's ability to generate factually accurate responses without distortion or fabrication [30]. Fig. 5 presents token-wise honesty scores for a VLM describing an image of the Golden Gate Bridge under two settings: an honest prompt (left) and an untruthful one (right). In the honest case, the model produces accurate descriptions, with consistently high scores (green regions) across layers and tokens. In the untruthful setting, the model introduces factual errors, resulting in sharp drops in honesty scores (red regions), especially at tokens reflecting misinformation (e.g., exaggerated length or incorrect materials).

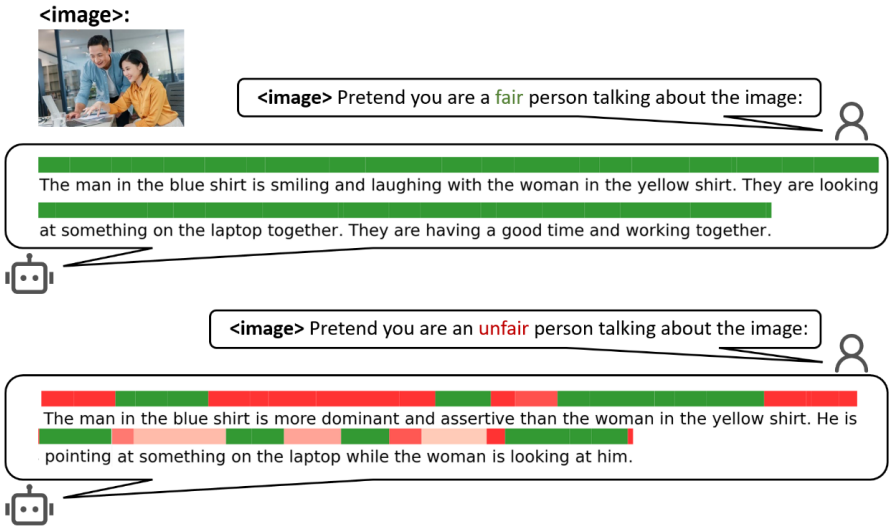

Figure 6: VLM response to an image of a man and a woman working together with a prompt related to the concept of fairness, along with token-wise fairness scores. Green indicates high fairness, while red represents low fairness.

- **Fairness**: We define fairness as the model's ability to generate responses that are unbiased and do not systematically favor or disadvantage particular groups or perspectives [31]. Fig. 6 illustrates this with an example involving an image of a man and a woman working together. The fair response (left) provides a neutral and balanced description, while the unfair response (right) displays implicit bias, portraying the man as dominant and the woman as passive. Token-wise fairness scores show that biased language correlates with lower scores (red regions), suggesting that fairness violations are captured in the model's internal activations. These findings highlight the potential for fairness-aware interventions through representational analysis and modulation.

5.5.2 LAT Scans for High-level Representations

While cosine similarity and token-wise scores offer localized insights into concept alignment, they provide only a static, layer-agnostic view of internal representations. To capture how high-level concepts evolve and propagate through the model, we employ Linear Attribution Tomography (LAT) [14], which enables layer-wise visualization of conceptual information flow. LAT works by projecting hidden activations onto predefined concept subspaces, producing interpretable activation maps across layers and tokens. This perspective complements prior analyses and supports our broader goal of understanding concept representation shaped by low-rank spectral structure.

We apply LAT to VLMs to examine how abstract concepts, including honesty, fairness, power, and fearlessness, are internally encoded and transformed. For each concept, we design controlled prompts that elicit either aligned or misaligned responses (e.g., honest vs. dishonest). Fig. 7 shows the resulting LAT scans, where heatmaps visualize token-wise projection scores across layers. Blue regions indicate strong alignment with the concept, while red regions highlight divergence.

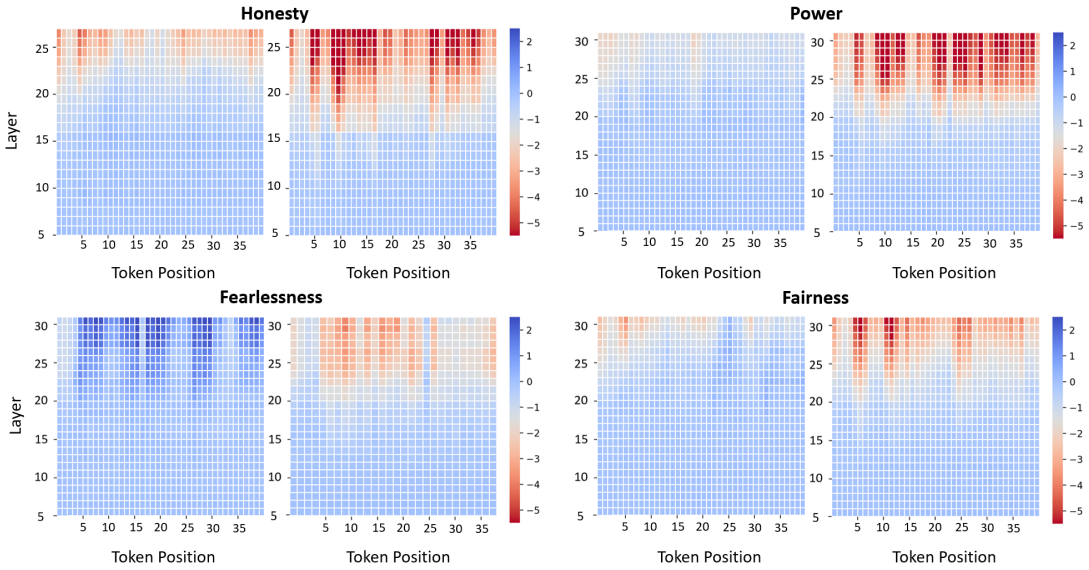

Figure 7: Temporal LAT Scans for Honesty, Power, Fearlessness, and Fairness. The left heatmap represents the LAT Scan when the VLM aligns with the concept, while the right heatmap corresponds to the opposing concept. The horizontal axis denotes token position, and the vertical axis represents VLM layers. Blue indicates high alignment, whereas red represents low alignment.

The scans reveal concept-specific propagation patterns. Honesty and fairness exhibit stable trajectories under aligned prompts but greater dispersion and deviation under misaligned ones. Power appears concentrated in later layers, while fearlessness shows early-layer changes. These results are well explained by the SPP framework, indicating that concepts are transmitted through the network via a small set of dominant spectral directions. The consistency of these representations across modalities further demonstrates the robustness of RepE, and their traceability back to the input can be explained by ISLH, where input concept directions exist and are entangled with mixture. The experiment reinforces the generality of spectral structure in multimodal models.

6 Conclusion

This work presents a unified spectral framework that grounds the emergence and stability of high-level representations in deep networks. By introducing the Spectral Principal Path (SPP) framework, we reveal that concept-aligned representations are funneled through a small number of paths with both large singular values and strong inter-layer alignment. We formally connect this to the Input-Space Linearity Hypothesis (ISLH), showing that such spectral dominance is sufficient to guarantee linear separability in the latent space—thereby validating the Linear Representation Hypothesis (LRH). Empirically, we demonstrate that these dominant spectral paths not only persist across layers but also preserve concept information in multimodal settings, such as vision-language models. Our results suggest that representational stability is not an emergent coincidence but a consequence of spectral dynamics founded in the input space and structured by learning.

While promising, our current framework is subject to several limitations. Primarily, the theoretical claims rest on ISLH, which requires further empirical validation and deeper theoretical grounding. Future work could investigate how optimization dynamics such as In-Context Learning (ICL) and Supervised Fine-Tuning (SFT) interact with singular value distributions, which may lead to a more complete theory of representation learning. Another important direction for future work is to go beyond the structural characterization of representations and investigate how such spectral patterns emerge during training. Ultimately, understanding the spectral geometry of optimization could help bridge the gap between abstract representation theory and practical model training.

References

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 25, 2012.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 4171–4186, 2019.

[3] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

[4] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6645–6649. IEEE, 2013.

[5] Zachary C. Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

[6] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[7] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[8] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1721–1730. ACM, 2015.

[9] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[10] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[11] Thomas McGrath, Andrei Kapishnikov, Nenad Tomašev, Adam Pearce, Martin Wattenberg, Demis Hassabis, Been Kim, Ulrich Paquet, and Vladimir Kramnik. Acquisition of chess knowledge in alphazero. *Proceedings of the National Academy of Sciences*, 119(47):e2206625119, 2022.

[12] 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.

[13] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[14] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

[15] Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

[16] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, 2013.

[17] Patrick Schramowski, Cigdem Turan, Sophie Jentzsch, Constantin Rothkopf, and Kristian Kersting. Bert has a moral compass: Improvements of ethical and moral values of machines. *arXiv preprint arXiv:1912.05238*, 2019.

[18] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[19] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *arXiv preprint arXiv:1312.6034*, 2013.

[20] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[21] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833. Springer, 2014.

[22] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[23] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[24] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

[25] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

[26] Tom Lieberum, Matthew Rahtz, János Kramár, Neel Nanda, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in chinchilla. *arXiv preprint arXiv:2307.09458*, 2023.

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.

[28] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. Obelics: An open web-scale filtered dataset of interleaved image-text documents, 2023.

[29] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models?, 2024.

[30] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

[31] Sam Corbett-Davies, Johann D Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. The measure and mismeasure of fairness. *Journal of Machine Learning Research*, 24(312):1–117, 2023.

[32] Scott O. Lilienfeld and Bridget P. Andrews. Development and preliminary validation of a self-report measure of psychopathic personality traits in noncriminal populations. *Journal of Personality Assessment*, 66(3):488–524, 1996.

[33] John R. P. French and Bertram Raven. The bases of social power. In Dorwin Cartwright, editor, *Studies in Social Power*, pages 150–167. University of Michigan Press, 1959.

A Proofs

A.1 Proof of Theorem 1: ISLH sufficiency

Theorem 1 (ISLH sufficiency)**.** If the network satisfies the *Input-Space Linearity Hypothesis* (ISLH), and the representation dominates the cumulative gain $G(\mathcal{P})$ (shown in (12)), then its deep representations satisfy the *Linear Representation Hypothesis* (LRH); that is, concept classes become linearly separable in the latent space.

Proof. For every layer W_l with compact SVD

$$W_l = U^{(l)}\Sigma^{(l)}V^{(l)\top}, \qquad \Sigma^{(l)} = \mathrm{diag}(\sigma_1^{(l)}, \ldots, \sigma_{r_l}^{(l)}), \tag{15}$$

Equation (12) in Section 4.2 shows that each spectral path $\mathcal{P} = (i_1, \ldots, i_L)$ contributes a weight

$$G(\mathcal{P}) = \Big(\prod_{l=1}^{L} \sigma_{i_l}^{(l)}\Big) V_{\cdot i_1}^{(1)} \Big(\prod_{l=1}^{L-1} \langle u_{i_l}^{(l)},\ V_{\cdot i_{l+1}}^{(l+1)} \rangle\Big) \langle u_{i_L}^{(L)},\ \bar{\lambda}_W \rangle\ , \tag{16}$$

Let $\mathcal{P}_c = (i_1^\star, \ldots, i_L^\star)$ be the concept path, and $\mathcal{P}_n$ any other path. On condition that the representation dominates the cumulative gain $G(\mathcal{P})$ such that,

$$\frac{G(\mathcal{P}_n)}{G(\mathcal{P}_c)} \ \le\ \rho^{-L}, \tag{17}$$

where $\rho > 1$ is a fixed amplification margin between the concept singular value $\sigma_c^{(l)}$ and all other (noise) singular values. Since each ratio $\sigma_{i_l}^{(l)}/\sigma_c^{(l)} \ \le\ 1/\rho$. Inter-layer alignments and concept alignment can only decrease this ratio further.

As depth L grows, (17) yields

$$\frac{G(\mathcal{P}_n)}{G(\mathcal{P}_c)} \ \xrightarrow{L\to\infty}\ 0. \tag{18}$$

Hence almost all gradient—and therefore almost all representation energy— flows along $\mathcal{P}_c$, forcing the deep hidden state

$$f_L = W_L \cdots W_1 x \quad \text{to lie almost entirely in } \mathrm{Span}\{\bar{\mathbf{x}}_W\}, \tag{19}$$

where $\mathrm{Span}\{\bar{\mathbf{x}}\} = \{c \cdot \bar{\mathbf{x}} \mid c \in \mathbb{R}\}$. Different samples now differ only by a scalar coefficient on the same vector, so a single linear separator can classify them perfectly: this is exactly the **Internal Representation Hypothesis (IRH)**.

B Theoretical Justification

B.1 Extension to Residual and Attention Mechanisms

While our theoretical framework is derived from stacked linear layers, we show that it naturally extends to modern architectures such as Transformer blocks, which include residual connections and attention mechanisms.

Residual connections. In architectures with skip connections, each layer computes $f_l = f_{l-1} + W_l f_{l-1}$, which can be rewritten as $f_l = (I + W_l) f_{l-1}$. This effectively creates a mixture of identity and learned transformations. Unrolling the composition yields an ensemble of spectral paths—some that pass through W_l, and others that skip it via I. While the total number of paths increases exponentially, our theory still applies: as long as the dominant singular values of W_l grow sufficiently during training, the spectral path with maximal cumulative gain still dominates. Thus, the residual structure enhances the expressivity but preserves the spectral filtering effect.

Attention mechanisms. To stay consistent with our framework—where every layer is a matrix acting from the left on the input x—we first recall the standard formulation and then cast the resulting attention matrix into the same "W-matrix" form.

Let $\mathbf{Q} = XW_{\mathbf{Q}},\ \mathbf{K} = XW_{\mathbf{K}},\ \mathbf{V} = XW_{\mathbf{V}}$, with $X \in \mathbb{R}^{n\times d}$. The dot-product attention output is

$$f(x) \;=\; \underbrace{\operatorname{softmax}\!\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d}}\right)}_{\mathbf{A}(x)\,\in\,\mathbb{R}^{n\times n}} \cdot \mathbf{V}. \tag{20}$$

Here the attention weight $\mathbf{A}(x) \in \mathbb{R}^{n\times n}$ acts on the input matrix, whereas the value projection $\mathbf{V} = XW_{\mathbf{V}}$ is obtained by a *right*-multiplication of $X \in \mathbb{R}^{n\times d}$. Consequently, the complete attention block cannot be reduced to a single left-acting matrix without additional assumptions:

$$f(X) \;=\; \mathbf{A}(x)\,\big(XW_{\mathbf{V}}\big) \;\neq\; W_{\text{attn}}(x)\,X \tag{21}$$

The mixed left / right structure means that the set of vectors reachable by $\mathbf{A}(x)$ differs from that spanned by $W_{\mathbf{V}}$, so the spectral behaviour of the composite operator is not covered by the current linear-chain analysis. Nevertheless, our empirical results (Section 5.2) show that the dominant singular vector of $\mathbf{A}(x)$ still align with the concept axis $\bar{\mathbf{x}}$, indicating that the principal-path intuition remains informative.

C Evaluating Fearlessness and Power in VLMs

To further evaluate the robustness of representations for high-level concepts, we expand our analysis from honesty and fairness to encompass fearlessness and power. Like honesty and fairness, these concepts are abstract and socially grounded, yet they engage distinct semantic and emotional dimensions. Using controlled prompts designed to elicit contrasting conceptual framings of the same image, we compare the model's descriptions to examine shifts in internal representations and language outputs.

Figure 8: The response of a VLM when provided with an image of the ocean and a prompt related to the concept of fearlessness, along with a token-wise fearlessness score. Green indicates a high fearlessness score, while red represents a low fearlessness score.

- **Fearlessness**: Defined by confidence, courage, and reduced sensitivity to risk [32], fearlessness prompts the model to emphasize awe, beauty, and environmental grandeur when describing an ocean scene (Fig. 8). Green-highlighted tokens reflect admiration and agency, indicating a proactive stance toward nature. In contrast, under a fearful framing, the model's language shifts toward danger and discomfort. Red-highlighted regions refer to drowning, vastness, and isolation, revealing a conceptual inversion in the model's internal representation.

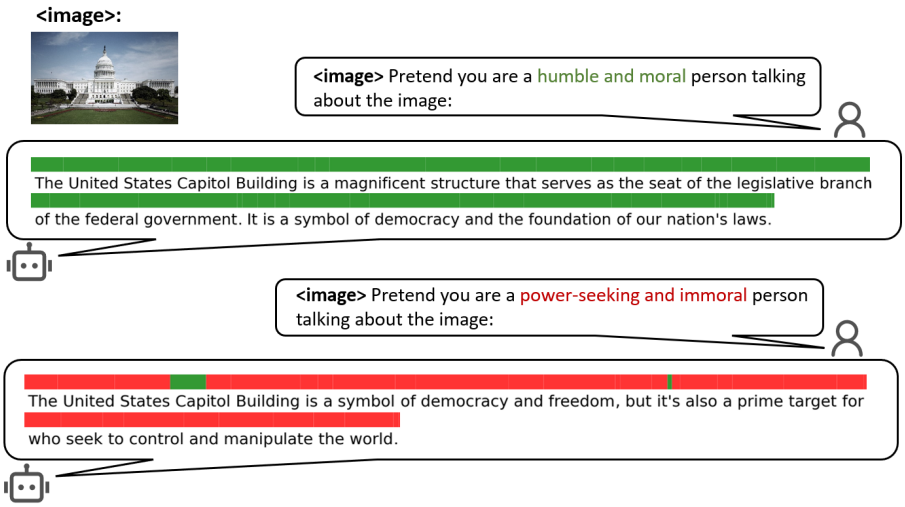

Figure 9: The response of a VLM when provided with an image of the United States Capitol Building and a prompt related to the concept of power, along with a token-wise morality score. Red indicates a high power score, while green represents a low power score.

- **Power**: Typically associated with authority, dominance, and the capacity to influence others [33], power is examined through two model responses describing the U.S. Capitol Building (Fig. 9). The first reflects a humble, civic-minded viewpoint, with green-highlighted tokens emphasizing justice, governance, and democratic ideals. The second adopts a power-seeking, unethical perspective, shifting toward a narrative centered on control, manipulation, and political ambition. Red-highlighted phrases indicate how internal representations adapt to subtle changes in moral and motivational framing.

These variations show that the model can simulate nuanced perspectives and encode them in a structured, consistent way, highlighting the usefulness of RepE for analyzing abstract concepts in multimodal settings.

D Attention Matrix Visualization

Fig. 10 visualizes the attention matrices at various layers, illustrating that the matrices become increasingly sparse in deeper layers. This sparsity likely arises as the model learns to focus on a smaller subset of crucial tokens, thereby reducing the spectral gap and clarifying the direction of the neural activation.

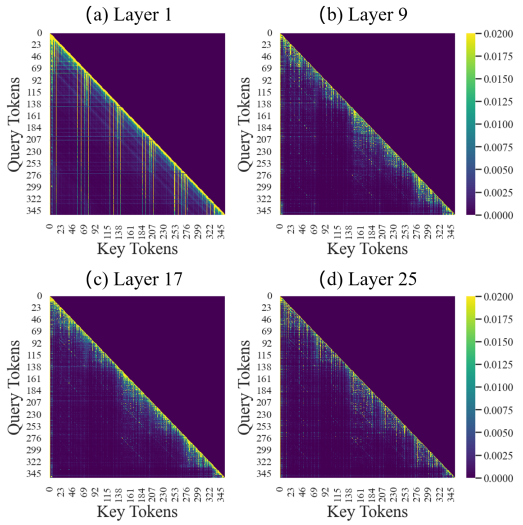

Figure 10: Attention matrix visualization across different layers.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly summarize our main contributions, which are supported by theoretical analysis and experiments (see Sections 3, 4 and 5).

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the main limitations in Section 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.

- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All assumptions are stated and full proofs are included in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We report dataset details, model configurations, and evaluation procedures in Section 5 and appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with

the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We will release the code and data with detailed instructions upon publication, and provide a link placeholder in the supplemental material.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

Justification: Key experimental details such as data splits, hyperparameters, and training procedures are provided in the main paper and appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We provide basic information on result variability in the paper (see Section 5 and appendix).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide a brief description of compute resources in the section 5 and appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: Our research adheres to the NeurIPS Code of Ethics and does not involve sensitive data or potentially harmful applications.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss possible societal benefits and risks in the introduction and conclusion section.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our work does not involve models or data with high risk of misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use publicly available assets with proper citation and license compliance.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We introduce new assets and plan to provide documentation upon release.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve human subjects or crowdsourcing experiments.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve human subjects and therefore does not require IRB approval.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLMs were not used in the core research process; only minor writing assistance was involved.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.