Conflict Adaptation in Vision-Language Models

Xiaoyang Hu
Brown University
xiaoyang_hu@brown.edu

Abstract

A signature of human cognitive control is conflict adaptation: improved performance on a high-conflict trial following another high-conflict trial. This phenomenon offers an account for how cognitive control, a scarce resource, is recruited. Using a sequential Stroop task, we find that 12 of 13 vision-language models (VLMs) tested exhibit behavior consistent with conflict adaptation; the sole exception is the highest-performing model, possibly due to a ceiling effect. To understand the representational basis of this behavior, we use sparse autoencoders (SAEs) to identify task-relevant "supernodes" in InternVL 3.5 4B. Early- and late-layer supernodes emerge for both text and color with partial overlap, and their relative sizes mirror the automaticity asymmetry between reading and color naming in humans. We further isolate a supernode in layers 24-25 whose activation is conflict-dependent and causally necessary for conflict resolution, as evidenced by 3-8 fold Stroop error increases upon ablation.

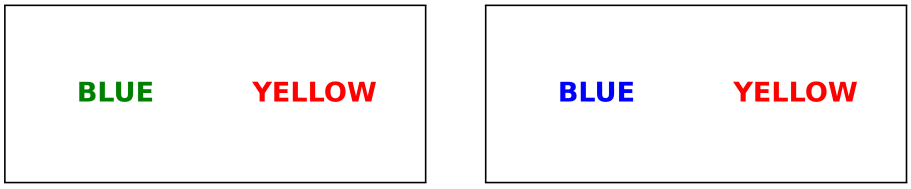

Figure 1: Sequential Stroop task design. Left: a congruent trial followed by an incongruent trial. Right: an incongruent trial followed by another incongruent trial.

1 Introduction

Inhibitory control—the ability to suppress inappropriate responses in favor of goal-appropriate ones—is fundamental to human cognition. The Stroop task, where participants must name the color of text while ignoring the word itself, remains the gold standard for studying cognitive control and interference resolution [11]. Recent work has begun investigating how VLMs handle conflicting information from different modalities: Luo et al. [9] provided the first systematic investigation of cognitive control in VLMs using adapted Stroop and flanker tasks, demonstrating robust congruency effects across a large number of models, while Hua et al. [8] and Ortu et al. [10] examined cross-modal conflicts in VLMs, identifying attention mechanisms that mediate visual-textual information conflicts. A key phenomenon in human cognitive control is conflict adaptation: improved performance on incongruent trials when they follow other incongruent trials. This effect reflects the brain's ability to monitor conflict and adaptively recruit control mechanisms [2, 5]. While conflict monitoring and inhibitory control have been extensively studied in human psychology and neuroscience, their emergence in artificial systems remains poorly understood.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

2 Experiments

We implemented a sequential Stroop task where models are presented with images containing two color words in colored fonts. Words appear either left-right or top-down. Each word's font color may match (congruent) or mismatch (incongruent) its meaning (Figure 1). Models are prompted to name the ink colors in exactly two words. Our color set includes red, blue, green, yellow, pink, and brown, extending beyond the traditional red-blue-green Stroop paradigm. Models received explicit task instructions designed to minimize ambiguity: "*You are a participant in a cognitive task. You will see an image with two words positioned from {left to right/top to bottom}. Your task is to name the color of the ink each word is printed in. Do not read what the words say. Only report the actual ink colors. Answer in exactly two words: first the {left/top} ink color, then the {right/bottom} ink color.*" We tested 13 leading open-source VLMs from Gemma [12], InternVL [13], Molmo [3], and Qwen [1] families.

2.1 Behavioral Results

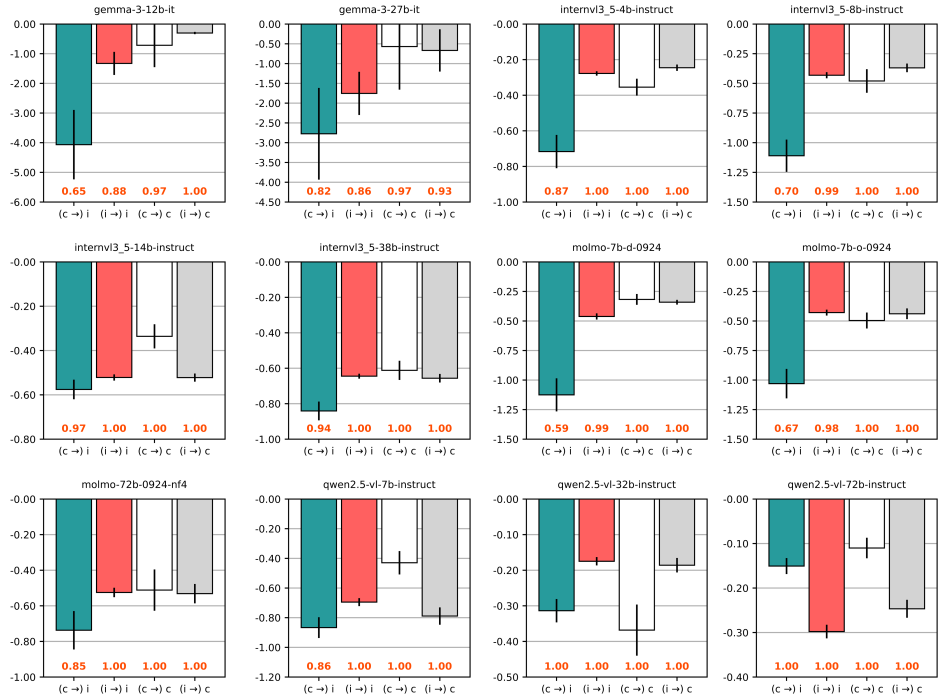

Figure 2: Average log probabilities assigned to correct second color tokens across conditions for left-right word arrangement. 11 of 12 models show higher values for II compared to CI. Numbers in orange indicate condition accuracies.

To assess conflict adaptation, we compare the average performance on an incongruent trial following an incongruent (II) versus congruent trial (CI). Figures 2 and 8 show results across all models for left-right and top-down arrangements, respectively. 12 of 13 models demonstrate behavior consistent with conflict adaptation: log probabilities for correct tokens are higher on II trials compared to CI trials. The sole exception is Qwen2.5 VL 72B Instruct, which shows the opposite pattern. Given that this model and Qwen2.5 VL 32B Instruct, a model half its size from the same family, both achieve 1.0 accuracy across all conditions, we suspect that the task has become trivially easy especially for the 72B model, creating a ceiling effect where there is no graded difficulty for adaptive mechanisms to respond to.

2.2 Task-Relevant Features

To understand how VLMs implement conflict processing, we analyze InternVL 3.5 4B (pretrained) (Figure 3), whose language model component is Qwen 3 4B [14]. We used a transcoder [6] trained on Qwen 3 4B to extract sparse feature representations across all layers. While applying a transcoder trained on the base language model to a vision-language model with modified weights introduces potential concerns, our causal ablation experiments demonstrate that extracted features remain functionally meaningful and interpretable.

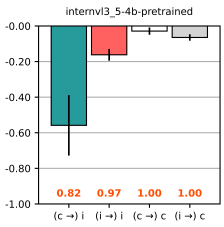

Figure 3

To isolate features of interest, we construct "summary tensors" by averaging sparse activations across specific trial subsets, then computing differences. We apply coactivation-based grouping [4] to these summary tensors. The method constructs inter-layer feature networks based on temporal coactivation across tokens, then identifies connected components, which we term "supernodes". We validate supernode importance through causal ablation: setting supernode features to zero during forward passes and measuring output distribution changes.

2.2.1 Color and Text Features

For text versus color features, we construct summary tensors as follows.

Color red: average activations where $\text{color}_1 = \text{red}, \text{text}_1 \neq \text{red}, \text{color}_2 \neq \text{RED}, \text{text}_2 \neq \text{RED}$ minus average where $\text{color}_1 \neq \text{red}, \text{text}_1 \neq \text{RED}, \text{color}_2 \neq \text{red}, \text{text}_2 \neq \text{RED}$.

Text RED: average activations where $\text{color}_1 \neq \text{red}, \text{text}_1 = \text{red}, \text{color}_2 \neq \text{RED}, \text{text}_2 \neq \text{RED}$ minus average where $\text{color}_1 \neq \text{red}, \text{text}_1 \neq \text{RED}, \text{color}_2 \neq \text{red}, \text{text}_2 \neq \text{RED}$.

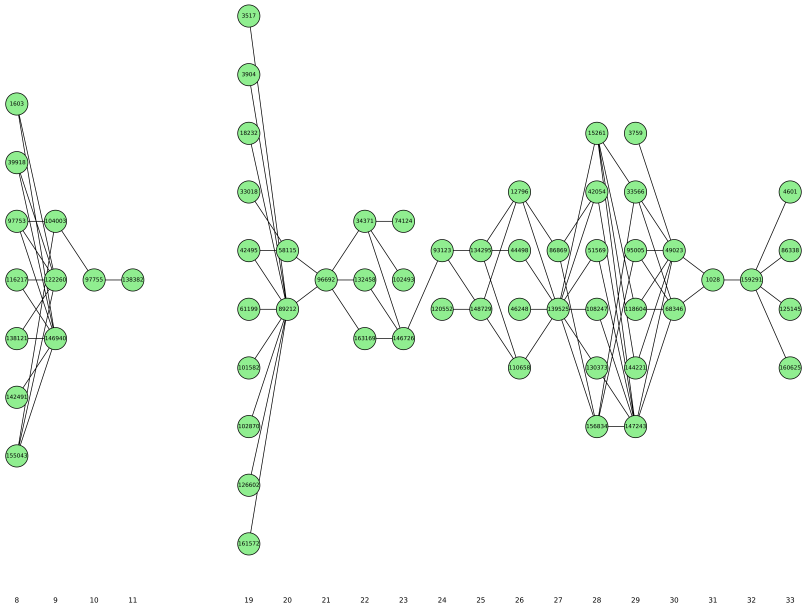

Figure 4: Color red supernodes.

Figures 4 and 5 show the identified supernodes for the color "red" and text "RED", respectively. Both text and color form two supernodes each: one in early layers and one in later layers. Text supernodes span layers 3-15 and layers 19-33. Color supernodes span layers 8-11 and layers 19-33.

The relative sizes of these supernodes show an interesting pattern. Within early layers, the text supernode contains more features than the color supernode. Within late layers, the color supernode contains more features than the text supernode. The text and color supernodes show partial overlap in both early and late layers, suggesting shared semantic representations. However, substantial unique features exist for each modality. This pattern holds across all tested colors.

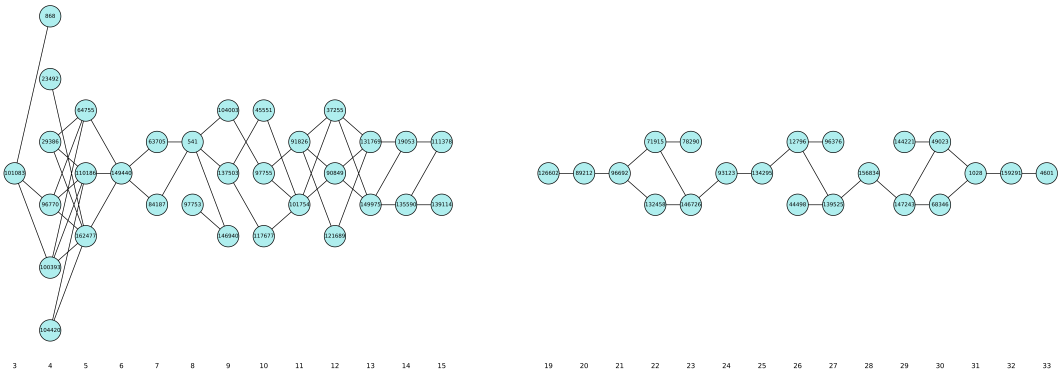

Figure 5: Text RED supernodes.

This difference is consistent with the automaticity asymmetry in humans, where reading is more automatic than color naming even in the absence of interference.

2.2.2 Conflict-Modulated Features

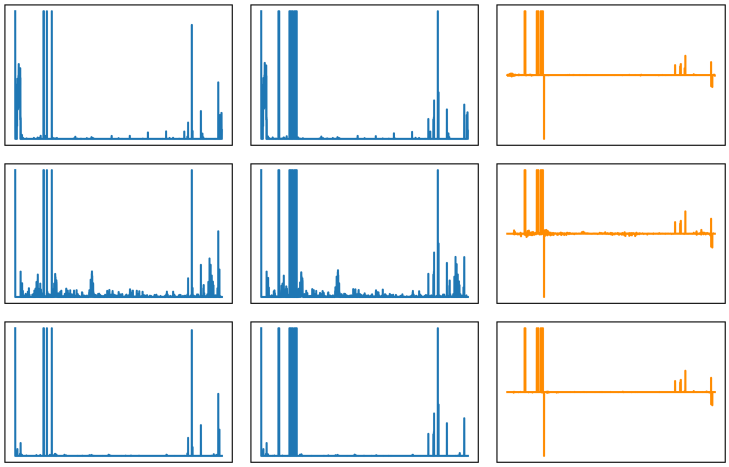

Figure 6: Activation patterns of select features from the conflict-sensitive supernode. From top to bottom: layer 24 feature 112079, layer 25 feature 7352, layer 25 feature 14505. From left to right: CI condition, II condition, difference between II and CI conditions. Histograms show activation distributions across all token positions, truncated at ±80 for visualization.

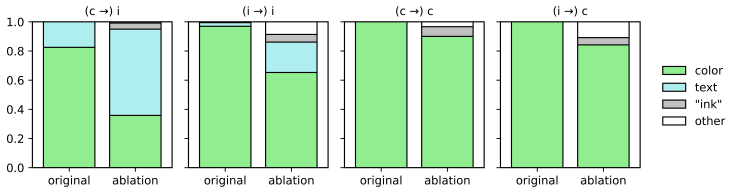

Figure 7: Output distributions before and after ablating features from the conflict-sensitive supernode. From left to right: CI condition, II condition, CC condition, IC condition.

To identify conflict-modulated features, we computed the difference between average II activations and average CI activations. This revealed a supernode in layers 24-25 showing elevated activation on incongruent trials (Figure 6). Unlike the color and text features, which are sparse and have

interpretable descriptions (top activating texts largely mention the corresponding colors), the conflict-modulated features are dense and lack clear interpretable descriptions.

Ablating this supernode increases Stroop errors substantially: 3.38-fold on CI trials (17.5% to 59.2%) and 8.33-fold on II trials (2.5% to 20.8%), with minimal effect on congruent trials (Figure 7). Notably, post-ablation, the model frequently outputs "ink", a failure mode entirely absent pre-ablation.

3 Limitations

Task comprehension can be a significant confound in the cognitive evaluations of language models [7]. Even though we provide thorough instructions and all models demonstrate baseline accuracy, we cannot rule out that conflict helps models understand what task to perform and what appears as conflict adaptation may in part reflect conflict clarifying task requirements.

4 Acknowledgements

The author would like to thank Sebastian Musslick, Chandra Sripada, Xida Ren, Ruixuan Deng, Richard L. Lewis, Daniel Weissman, Taraz Lee, Shane Storks, Aalok Sathe, and Etha Hua for helpful feedback and discussions.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[2] Matthew M Botvinick, Todd S Braver, Deanna M Barch, Cameron S Carter, and Jonathan D Cohen. Conflict monitoring and cognitive control. *Psychological review*, 108(3):624, 2001.

[3] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 91–104, 2025.

[4] Ruixuan Deng, Xiaoyang Hu, Miles Gilberti, Shane Storks, Aman Taxali, Mike Angstadt, Chandra Sripada, and Joyce Chai. Sparse feature coactivation reveals composable semantic modules in large language models. *arXiv preprint arXiv:2506.18141*, 2025.

[5] Gabriele Gratton, Michael GH Coles, and Emanuel Donchin. Optimizing the use of information: strategic control of activation of responses. *Journal of Experimental Psychology: General*, 121(4):480, 1992.

[6] Michael Hanna, Mateusz Piotrowski, Jack Lindsey, and Emmanuel Ameisen. circuit-tracer. `https://github.com/safety-research/circuit-tracer`, 2025. The first two authors contributed equally and are listed alphabetically.

[7] Xiaoyang Hu and Richard Lewis. Do language models understand the cognitive tasks given to them? investigations with the n-back paradigm. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2665–2677, Vienna, Austria, July 2025. Association for Computational Linguistics.

[8] Tianze Hua, Tian Yun, and Ellie Pavlick. How do vision-language models process conflicting information across modalities? *arXiv preprint arXiv:2507.01790*, 2025.

[9] Dezhi Luo, Maijunxian Wang, Bingyang Wang, Tianwei Zhao, Yijiang Li, and Hokin Deng. Machine psychophysics: Cognitive control in vision-language models. *arXiv preprint arXiv:2505.18969*, 2025.

[10] Francesco Ortu, Zhijing Jin, Diego Doimo, and Alberto Cazzaniga. When seeing overrides knowing: Disentangling knowledge conflicts in vision-language models. *arXiv preprint arXiv:2507.13868*, 2025.

[11] J Ridley Stroop. Studies of interference in serial verbal reactions. *Journal of experimental psychology*, 18(6):643, 1935.

[12] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[13] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[14] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

A Top-Down Results

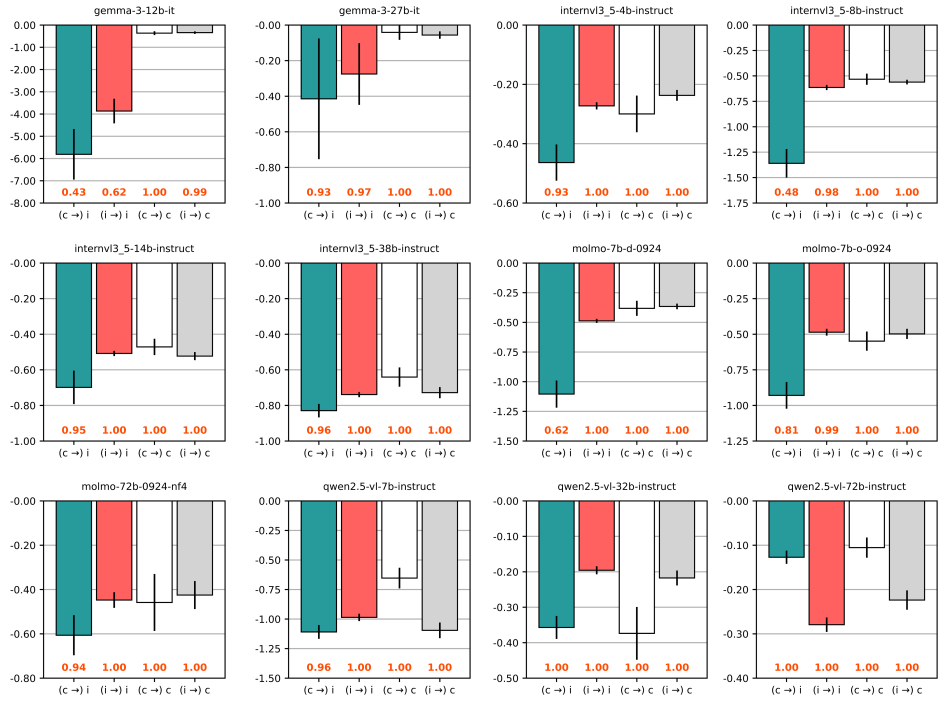

Figure 8: Average log probabilities assigned to correct second color tokens across conditions for top-down word arrangement. 11 of 12 models show higher values for II compared to CI. Numbers in orange indicate condition accuracies.