Machine Psychophysics: Cognitive Control in Vision-Language Models

Dezhi Luo[†] **(ihzedoul@umich.edu)**
University of Michigan

Maijunxian Wang (mjxwang@ucdavis.edu)
University of California, Davis

Bingyang Wang[*] **(icy.bingyang.wang@alumni.emory.edu)**
Emory University

Tianwei Zhao[*] **(tzhao27@jh.edu)**
Johns Hopkins University

Yijiang Li (yijiangli@ucsd.edu)
University of California San Diego

Hokin Deng[†] **(hokind@andrew.cmu.edu)**
Carnegie Mellon University

[*] Equal Contributions.

[†] Correspondence.

Abstract

Cognitive control refers to the ability to flexibly coordinate thought and action in pursuit of internal goals. A standard method for assessing cognitive control involves conflict tasks that contrast congruent and incongruent trials, measuring the ability to prioritize relevant information while suppressing interference. We evaluated 108 vision-language models on three classic conflict tasks and their more demanding ”squared” variants across 2,220 trials. Model performance corresponds closely to human behavior under resource constraints and reveals individual differences. This results indicate that some form of human-like executive function—albeit limited—may have emerged in current multi-modal foundational models.

Keywords: vision-language models; cognitive control; conflict processing; representational alignment; general intelligence

Introduction

Human behavior is distinguished by its flexibility and goal-directedness: we can pursue novel, underspecified tasks, adapt to changing contexts, and manage competing objectives over time (Rasmussen, 1990; Botvinick et al., 2001). At the core of these abilities lies cognitive control, a set of mechanisms that support the dynamic coordination of thought and action in service of internal goals (Egner, 2023; Badre, 2024), making it a particularly valuable target for evaluating signatures and guiding the development of prospective general intelligence in artificial systems (Anderson, 1983; Russin et al., 2020; LeCun, 2022).

A standard approach to assessing cognitive control is through tasks that elicit cognitive conflict, typically by contrasting congruent and incongruent trials. In these tasks, participants must respond to task-relevant features of a stimulus while ignoring distracting or conflicting ones (MacLeod, 1991). Performance differences between congruent and incongruent trials provide a window into the ability to maintain focus and suppress interference—capacities central to flexible, goal-directed behavior.

Vision-language models (VLMs) can integrate visual and textual information and have demonstrated strong performance on high-level reasoning benchmarks. Here, we evaluate 108 models on three classic conflict tasks, along with their more cognitively demanding ”squared” versions, in a large-scale, strictly controlled setting spanning 2,220 trials. Model performance corresponds remarkably to human behavior under limited computational resources and reveals robust individual differences, suggesting emergent cognitive mechanisms for executive functions, though overall competence remains limited in zero-shot, unconstrained settings.

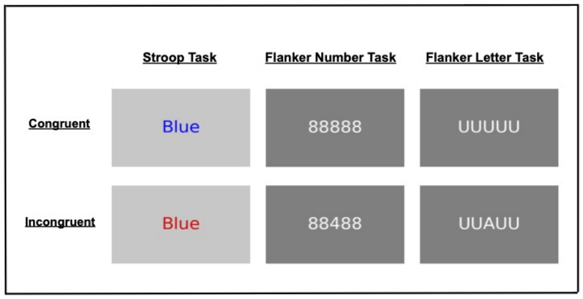

Figure 1: **Standard Tasks**. In the Stroop task, participants must indicate the color a word is printed in while disregarding the word’s meaning. In the Flanker tasks, participants must identify either the central letter or number while ignoring the surrounding distractors.

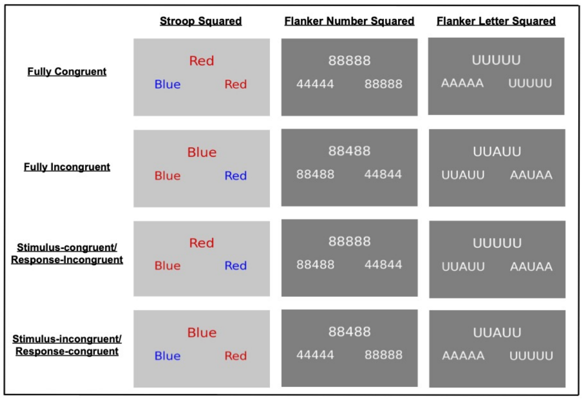

Figure 2: **Squared Tasks**. In Stroop Squared, participants select the response option whose word meaning matches the display color of the target word. In Flanker Squared, they choose the option where the central letter or number matches the identity of the surrounding distractors in the target stimulus. The correct response for all example trials shown is the option on the right.

Methods

We applied classic cognitive control tasks to evaluate models’ ability to resolve cognitive conflict. Specifically, we implemented the Stroop task (Stroop, 1935) and both the Letter and Number versions of the Flanker task (Eriksen & Eriksen, 1974). In parallel, we adapted the ”squared” design introduced by Burgoyne et al. (2023) across all three task types. This design adds an additional layer of conflict by requiring subjects to choose among response options that are congruent or incongruent with the target along dimensions already manipulated in the standard versions of the tasks.

We adopted the squared design for two reasons. First, classic conflict tasks often show limited between-participant variability, reducing their reliability for assessing individual differences (Hedge et al., 2018). In contrast, squared paradigms

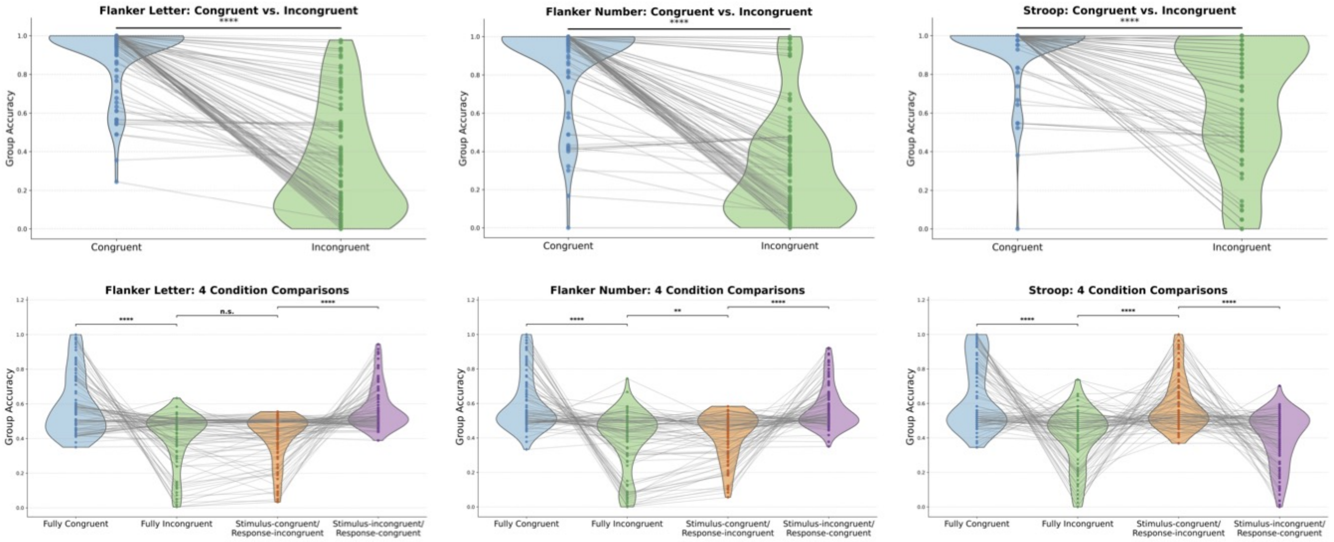

Figure 3: **Model Performances on Standard and Squared Tasks Compared Between Conditions.** Across all standard tasks, VLMs showed strong congruency effects, with significantly higher accuracy on congruent than incongruent trials: Flanker Letter ($t = 17.88$, $p < 10^{-33}$), Flanker Number ($t = 16.85$, $p < 10^{-31}$), and Stroop ($t = 8.99$, $p < 10^{-14}$). This pattern extends to the squared task variants within the comparison between FC and FI trials. VLMs displayed differentiated performance across all four conditions, with significant contrasts found in nearly all pairwise comparisons ($p < .001$), except in Flanker Letter, where FI and SCRI did not differ significantly ($t = -0.86$, $p = .39$).

elicit substantially greater variability across participants. Second, the increased complexity of squared tasks boosts cognitive demand, which is especially valuable when evaluating models that perform at ceiling on standard tasks—whether due to genuine competence or confounding heuristics. For example, a model biased toward color processing might bypass interference in the standard Stroop task, yet still struggle with the hierarchical conflict in the squared variant.

For the Stroop task, we selected 7 commonly identifiable colors and crossed them following both the standard and squared designs, yielding 84 images for the standard version (42 congruent [C], 42 incongruent [I]) and 336 for the squared version (84 each for fully congruent [FC], fully incongruent [FI], stimulus-congruent/response-incongruent [SCRI], and stimulus-incongruent/response-congruent [SIRC] conditions). For the two Flanker tasks (Letter and Number), we used all 10 single-digit Arabic numerals and 10 randomly selected letters from the English alphabet, generating 180 images for each standard task (90 congruent, 90 incongruent) and 720 for each squared version (180 per condition: FC, FI, SCRI, SIRC). All stimuli were paired with task-specific prompts in a binary forced-choice format, with response options counterbalanced within each task type.

Results

Across all standard tasks, VLMs showed robust congruency effects, with significantly higher accuracy on congruent than incongruent trials. This pattern extended to the squared tasks, where models exhibited clear performance differences across conflict conditions.

Moreover, robust individual differences emerged across all tasks, with distinctions becoming more pronounced under increased task difficulty. Standard tasks separated models into at-chance, conflicted, and near-perfect performers, while squared tasks further differentiated models by exposing persistent conflict sensitivity—even among high-performing models like GPT-4o, which resolved standard conflicts but remained vulnerable under hierarchical interference. These patterns closely resemble those observed in humans under constrained cognitive resources, as indexed by processing time (Lee et al., 2025)[1]. Our data suggest that models engage a general cognitive control mechanism across tasks, with some models exhibiting substantially more flexible and consistent control than others

Conclusion

We showed that model performance closely aligns with human behavior under resource constraints and reveals robust individual differences. These results provide substantial support for the emergence of a limited form of human-like executive function in current multimodal foundation models.

Acknowledgment

We thank Zillion Network Inc. for providing the computation used in this work. Their optimized peak/off-peak scheduling,

[1] For further details, we refer to the full version of the paper (Luo et al., 2025).

high-throughput storage infrastructure, and automated environment management enabled the cost-efficient and reliable execution of our experiments. We also thank the anonymous reviewers for their generous feedback, which are addressed in the full version.

References

Anderson, J. R. (1983). *The architecture of cognition*. Psychology Press.

Badre, D. (2024). Cognitive control. *Annual Review of Psychology*, *76*.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological review*, *108*(3), 624.

Burgoyne, A. P., Tsukahara, J. S., Mashburn, C. A., Pak, R., & Engle, R. W. (2023). Nature and measurement of attention control. *Journal of Experimental Psychology: General*, *152*(8), 2369.

Egner, T. (2023). Principles of cognitive control over task focus and task switching. *Nature Reviews Psychology*, *2*(11), 702–714.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & psychophysics*, *16*(1), 143–149.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior research methods*, *50*, 1166–1186.

LeCun, Y. (2022). A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, *62*(1), 1–62.

Lee, T. G., Sellers, J., Jonides, J., & Zhang, H. (2025). The forced-response method: A new chronometric approach to measure conflict processing. *Behavior Research Methods*, *57*(1), 1–14.

Luo, D., Wang, M., Wang, B., Zhao, T., Li, Y., & Deng, H. (2025). Machine psychophysics: Cognitive control in vision-language models. *arXiv preprint arXiv:2505.18969*.

MacLeod, C. M. (1991). Half a century of research on the stroop effect: an integrative review. *Psychological bulletin*, *109*(2), 163.

Rasmussen, J. (1990). The role of error in organizing behaviour. *Ergonomics*, *33*(10-11), 1185–1199.

Russin, J., O'Reilly, R. C., & Bengio, Y. (2020). Deep learning needs a prefrontal cortex. *"Bridging AI and Cognitive Science" Workshop at the International Conference on Learning Representation (ICLR)*, *107*(603-616), 1.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of experimental psychology*, *18*(6), 643.