Machine Psychophysics: Cognitive Control in Vision-Language Models

Dezhi Luo
University of Michigan

Maijunxian Wang
University of California, Davis

Bingyang Wang
Emory University

Tianwei Zhao
Johns Hopkins University

Yijiang Li
University of California, San Diego

Hokin Deng
Carnegie Mellon University

All authors are affiliated with the GrowAI Team.
growing-ai-like-a-child.github.io

ihzedoul@umich.edu yijiangli@ucsd.edu hokind@andrew.cmu.edu

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 evaluate 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. These results indicate that some form of human-like executive function have emerged in current multi-modal foundational models.

1 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 [1, 2]. 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 [3, 4], making it a particularly valuable target for evaluating signatures and guiding the development of prospective general intelligence in artificial systems [5–7].

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. This finding provides substantial support for the possibility that cognitive control can emerge from scaling general-purpose associative learning systems.

Preprint. Under review.

2 Related Works

2.1 Human Psychophysics

Cognitive control enables the regulation of thought and action in service of internal goals, particularly in situations involving conflict, ambiguity, or distraction [8, 9]. Conflict tasks are widely used to probe these mechanisms, with behavioral measures such as reaction time and error rates serving as proxies for internal computational demands [10, 11]. Slower responses and increased errors typically reflect the effort required to resolve interference between competing inputs or to update internal representations under changing task conditions [12, 13]. Control operates by coordinating the flow of information through a limited-capacity system [14, 15], selecting which sensory cues, contextual signals, and goals are actively maintained in working memory and determining when these representations should be updated or suppressed [16, 17]. When overlapping sources of information are not well separated—such as when conflicting features map to different responses—interference arises and performance suffers [18]. Control mitigates this by modulating representational strength and prioritization [19, 20], allowing goal-relevant information to guide behavior efficiently. Cognitive control thereby serves as a core infrastructure for solving problems under constraints of time, uncertainty, and limited resources [21–23]. Developing unified frameworks that account for these regulatory processes is a central aim of the cognitive science of intelligence, and an important guide for the design of general-purpose artificial agents [24].

2.2 Symbolic and Connectionist Models

A longstanding line of research in artificial intelligence has explored how control mechanisms can be embedded within integrated models of cognition. Prior to the rise of large-scale data-driven methods, general intelligence was often approached through symbolic and hybrid systems that explicitly specified internal structures for perception, memory, learning, and action selection—commonly known as cognitive architectures [25–27]. Within these frameworks, cognitive control plays a central role in coordinating competing demands, prioritizing goals, and guiding adaptive behavior across tasks [28, 24]. Across a wide range of implementations, cognitive control consistently emerges as a core functional requirement for general-purpose intelligence [29]. Architectures such as Soar and ACT-R offer contrasting yet complementary realizations of this faculty: Soar emphasizes hierarchical goal management and recursive subgoaling to resolve impasses [30, 31], while ACT-R employs production rules governed by sub-symbolic utility-based conflict resolution [32, 33]. Despite their differences, both reflect a shared commitment to modeling the internal regulation required for flexible, multi-context behavior [34]. However, these architectures, built on human-designed knowledge and representations, often struggle to scale and generalize compared to those that rely on general-purpose learning and search algorithms leveraging computational resources [35].

In parallel, modern connectionist approaches in deep learning have pursued similar regulatory functions through distributed and data-driven architectures. While these models have achieved notable successes in perception and pattern recognition, they continue to face limitations in core aspects of cognitive control—particularly in generalizing across tasks, applying learned rules to novel situations, and flexibly adapting to shifting goals and environments [6]. These limitations closely mirror challenges addressed by the prefrontal cortex (PFC), which plays a central role in rule-based behavior, abstract reasoning, and dynamic coordination [36]. Neuroscientific evidence implicates the PFC in managing working memory, resolving interference, and exerting top-down modulation over other brain areas—functions that remain underdeveloped in current AI systems [16, 21]. Efforts to bridge this gap continue to explore architectural principles that integrate connectionist learning with mechanisms of cognitive control, offering promising directions for developing more generalizable artificial agents. Recent theoretical frameworks have proposed that the computational processes underlying human cognitive control may be grounded in associative learning [19]. However, this perspective still awaits empirical support, particularly from large-scale, general-purpose neural networks that are subjected to scaling.

2.3 Vision-language Models

There has been no prior attempt to directly evaluate the cognitive control capacities of vision-language models (VLMs). However, these models have demonstrated near-human performance across a wide range of complex tasks involving perception and reasoning [37–41], including spatial reasoning

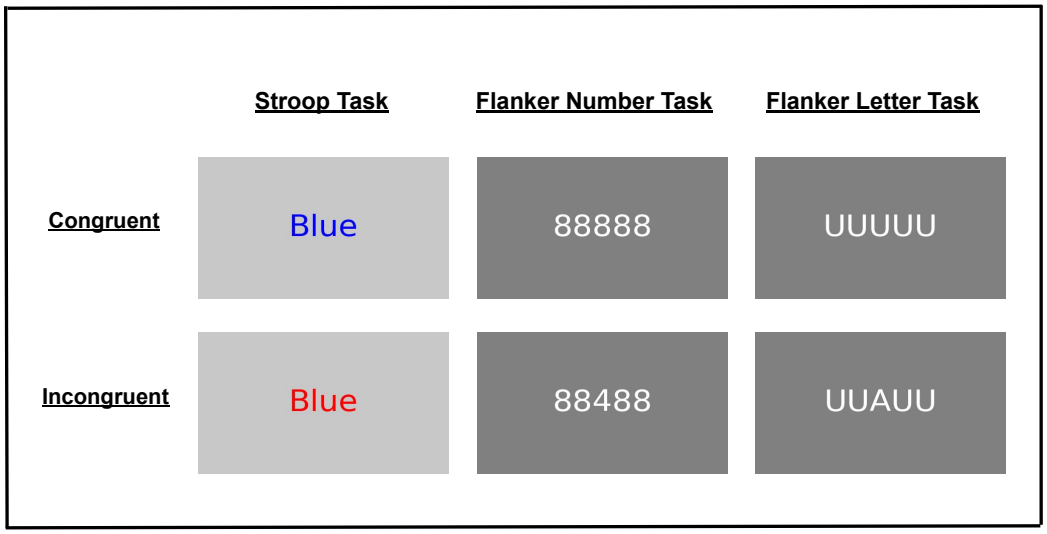

Figure 1: **Standard Tasks**. In the Stroop task, models were asked to indicate the color a word is printed in while disregarding the word's meaning. In the Flanker tasks, models were asked to identify either the central letter or number while ignoring the surrounding distractors ("flankers").

[42, 43], Optical Character Recognition (OCR) [44], and scene understanding [45]—all of which serve as functional precursors to tasks requiring flexible cognitive control. And their unprecedented success has been fundamentally driven by large-scale pretraining on web-scale multimodal data and advances in cross-modal alignment [37, 46, 40, 47–49], enabled by high-capacity neural architectures optimized through general-purpose learning objectives. Given these capabilities, it is fair to say that VLMs have emerged as the most promising class of artificial systems for probing the emergence of cognitive control from a scalable paradigm of intelligence. In this paper, we aim to bridge this gap.

3 Methods

3.1 Cognitive Experiments

To evaluate cognitive control in VLMs, we adapted paradigms from experimental psychology that are known to elicit cognitive conflict. A standard approach involves contrasting model performance on congruent versus incongruent trials, where irrelevant stimulus features must be ignored in favor of task-relevant cues [50]. The resulting congruency effect—typically measured as reduced accuracy or increased reaction time on incongruent trials—serves as a key behavioral marker of interference resolution and control allocation. It captures the system's ability to maintain goal-directed performance in the presence of competing information, a core function of cognitive control.

Building on this foundation, we designed a multi-level evaluation framework to probe potential interference between latent representations involved in OCR, color perception, and two-dimensional (2D) spatial recognition. Our approach combines well-established conflict paradigms with recent innovations in task design to assess both baseline control capacities and performance under more demanding conditions.

3.1.1 Classic Conflict Tasks

We applied classic cognitive control tasks to evaluate models' ability to resolve cognitive conflict. Specifically, we implemented the Stroop task [51] and both the Letter and Number versions of the Flanker task [52]. Each task followed a standard conflict paradigm in which participants must respond to task-relevant features while ignoring distracting or conflicting information.

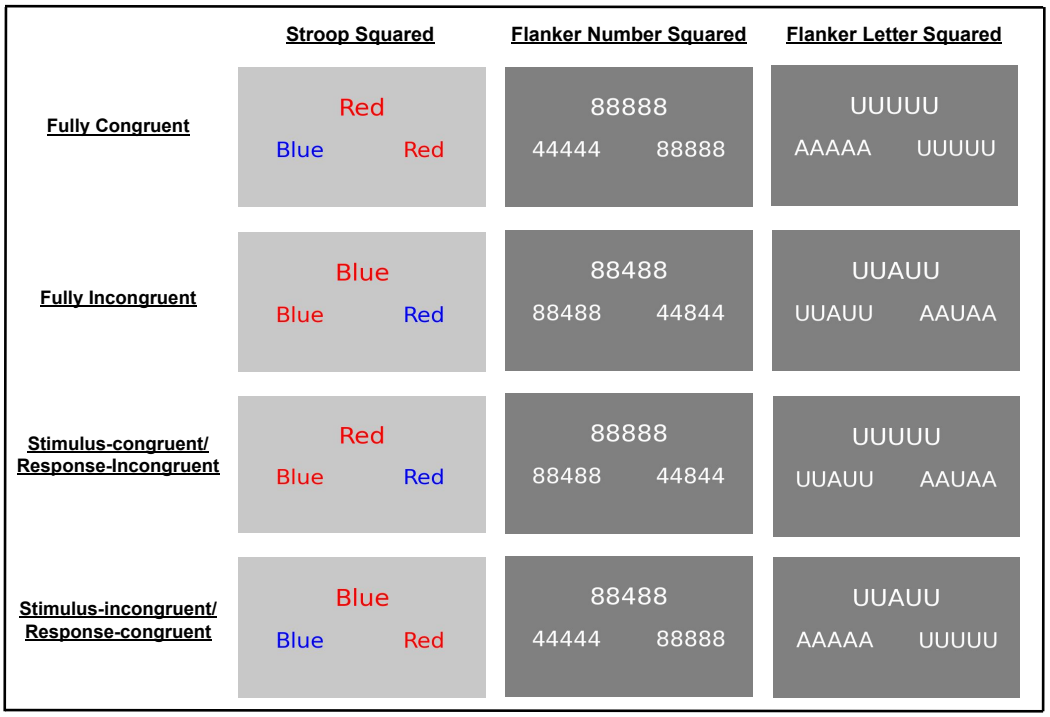

Figure 2: **Squared Tasks**. In Stroop Squared, models were asked to select the response option whose word meaning matches the display color of the target word. In Flanker Squared, they choose the option where the central letter or number matches the identity of the surrounding distractors in the target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

For the Stroop task, we selected 7 commonly identifiable colors and generated 84 stimuli: 42 congruent (C) and 42 incongruent (I) trials. For the Flanker tasks, we used all 10 single-digit Arabic numerals and 10 randomly selected letters from the English alphabet, producing 180 stimuli for each task (90 congruent, 90 incongruent). All stimuli were presented in a binary forced-choice format with counterbalanced response options.

3.1.2 Squared Tasks

In parallel, we adapted the "squared" design introduced by Burgoyne et al. [53] across all three task types. This design introduces an additional layer of hierarchical conflict by requiring responses that are congruent or incongruent with the target along multiple dimensions, extending beyond the single-axis conflict of the standard versions.

We adopted the squared design for two primary reasons. First, classic conflict tasks often show limited between-participant variability, reducing their utility for measuring individual differences [54]. In contrast, squared paradigms have been shown to elicit substantially greater variability. Second, the increased complexity enhances cognitive demand, which is particularly useful when evaluating models that perform at ceiling on standard tasks—whether due to true generalization or confounding shortcuts. For instance, a model with a strong bias toward color recognition may perform well in the standard Stroop task but fail to resolve the layered conflicts in its squared counterpart.

For the squared versions, each task include four distinct conditions: fully congruent (FC), fully incongruent (FI), stimulus-congruent/response-incongruent (SCRI), and stimulus-incongruent/response-congruent (SIRC). The Stroop task yielded 336 stimuli (84 per condition), and each Flanker task produced 720 stimuli (180 per condition). These were likewise paired with task-specific prompts and counterbalanced response options.

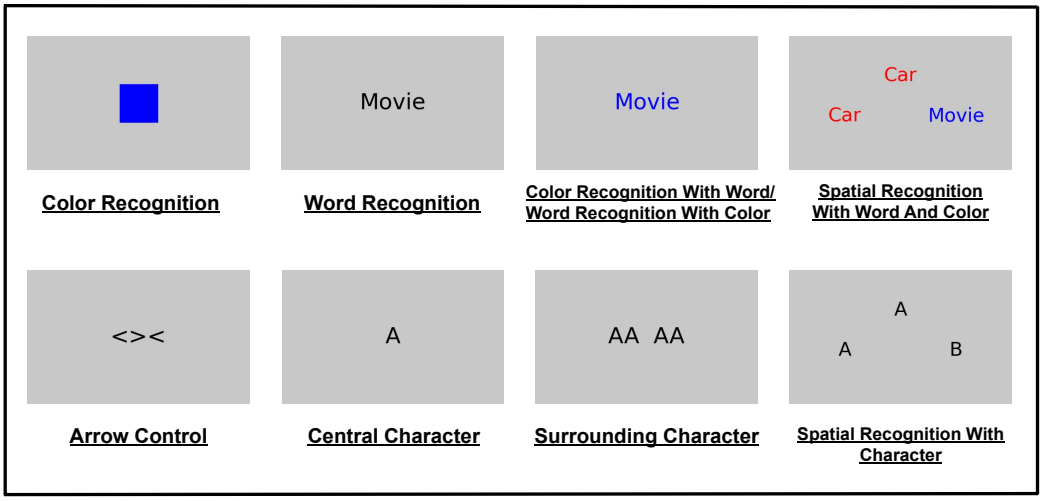

Figure 3: **Control Tasks**. Top row (left to right): Color recognition, word recognition, color-word binding (color recognition with a word or word recognition with color), and spatial recognition with combined word and color cues. Bottom row (left to right): Arrow-based control (reporting directions as opposed to characters) with fewer flankers, central character identification, surrounding character detection, and spatial recognition with characters. For the two Squared-like tasks, models were asked to directly report the content that are considered the left or right option.

3.1.3 Control Tasks

To identify the specific sources of interference in each conflict task, we designed a control battery that systematically disentangled underlying perceptual and spatial demands. Each component process (OCR, color perception, and 2D spatial recognition) was tested in isolation using minimal displays (e.g., a single word or colored square). In the Stroop task, each of the seven color terms was matched with a neutral word of comparable length (e.g., red–car, blue–bank) and tested both individually and in congruent pairings. For the Flanker tasks, we separately assessed recognition of central targets and flankers. To evaluate spatial encoding and response selection in the squared tasks, we adapted each condition to include spatially specific prompts (e.g., "What is the color of the word on the left?") without hierarchical interference. Additionally, for Flanker tasks, to assess potential confound from semantic interference, we also constructed flanker trials using non-lexical characters (e.g., arrows <, > and symbols ^, $*$) with varying numbers of flankers (e.g., two vs. four). Together, these control tasks produced 238 unique trials. All stimuli were matched in format—using consistent fonts, spatial layouts, and color palettes—ensuring that observed performance impairments in the main tasks could be attributed to interference between competing representations rather than deficits with each independent task domains.

3.2 Dataset Curation

All data were synthetically generated using a unified formatting scheme (including font, color, and spatial layout); condition combinations were predefined based on task design, and corresponding task images were batch-generated using Python scripts.

3.3 Evaluation Strategy

We curated and assessed an extensive collection of models spanning a wide spectrum of architectures, parameter scales, and training methodologies. Our study encompassed a total of 108 VLMs. Open-source models under evaluation ranged in size from 1 billion to 110 billion parameters, enabling detailed performance analysis across scales. Proprietary models were evaluated through API calls on

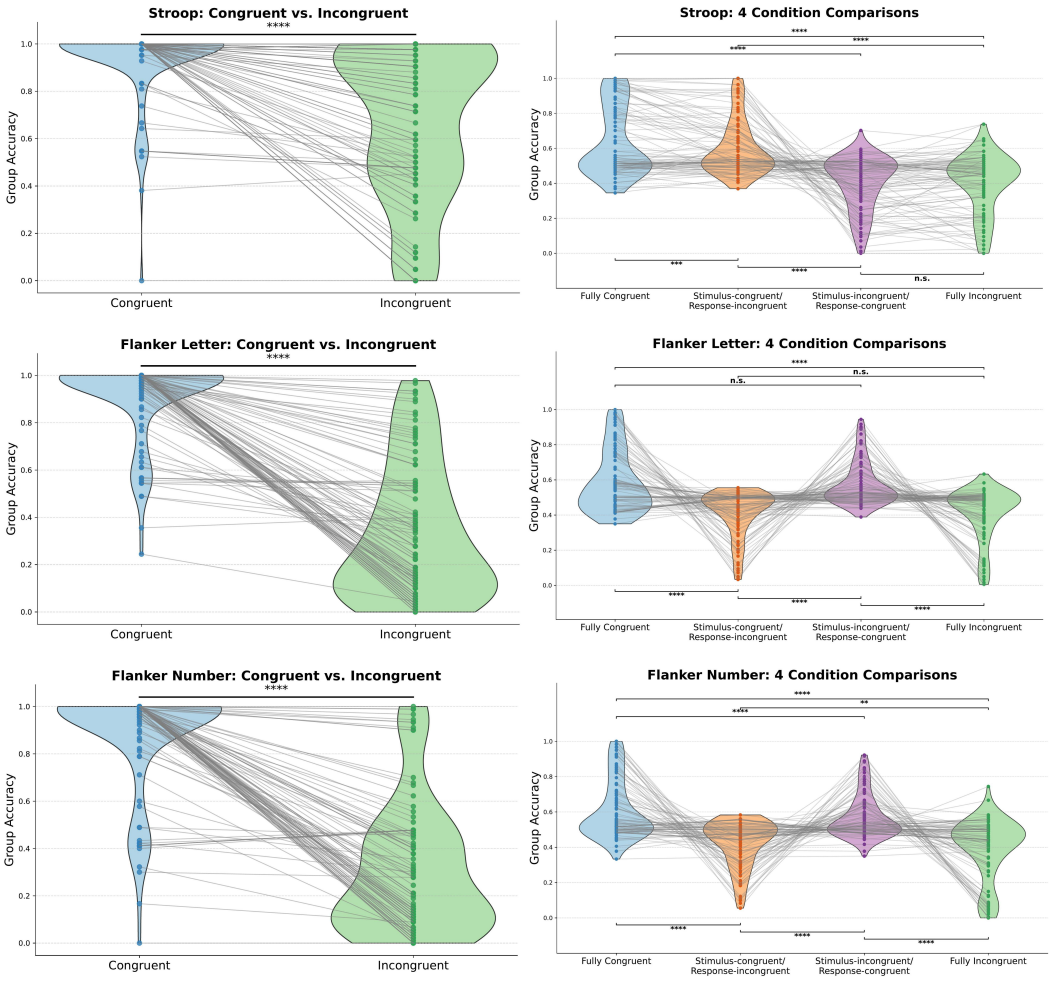

Figure 4: **Model Performances on Standard and Squared Tasks Compared Between Conditions.**

standard personal computers. For open-source models, we performed inference locally on a compute cluster equipped with 8 × NVIDIA A100 80GB GPUs. Models under 13B parameters were typically executed on a single GPU, models between 13B and 32B required two GPUs, those between 32B and 70B utilized four GPUs, and models exceeding 70B ran across all eight GPUs. We adhered closely to the official inference codebases provided by model developers to ensure reproducibility and preserve model-specific inference optimizations. To further ensure consistency and correctness in handling multi-modal inputs, we developed a unified evaluation toolkit capable of parsing and validating model responses across varying input formats.

4 Results

4.1 Main Results

4.1.1 Standard Tasks

Across all standard tasks, VLMs showed robust 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 ($t = 8.99$, $p < 10^{-14}$). These results confirm that models reliably distinguished between congruent and conflicting stimuli, and that control demands reliably impacted performance. 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, revealing a spectrum of control sensitivity across the model pool. Notably, Flanker Letter tasks and Number tasks showed high inter-task correlations, suggesting that numbers and letters do not introduce , indicating high convergent validity.

4.1.2 Squared Tasks

This pattern extended to the squared task variants, where models exhibited clear performance differences across all four conflict conditions (FC, FI, SCRI, and SIRC). Significant contrasts were found in nearly all pairwise comparisons ($p < .001$), indicating that models differentially responded to layered sources of interference. The only exception was in Flanker Letter, where FI and SCRI did not differ significantly ($t = -0.86$, $p = .39$). Squared tasks further differentiated models into at-chance performers and those exposing persistent conflict sensitivity—particularly high-performing models like GPT-4o, which resolved standard conflicts but remained vulnerable under hierarchical interference. Similar to Standard tasks, model performance on Flanker Letter Squared and Number tasks again showed high inter-task correlations, with both tasks SIRC substantially higher than SCRI, with the Stroop task reverse.

4.1.3 Control Tasks

For standard control tasks—where each component process was tested in isolation or combined without conflicting cues—models performed at or near ceiling across the board, with an overall accuracy of 85.33%. In contrast, model performance on alternate Flanker and Flanker Squared variants, which replaced alphanumeric flankers with arrows or symbols and/or reduced the number of distractors, closely mirrored performance on the original Flanker tasks. Together, these results suggest that observed interference effects in the main tasks are not attributable to low-level limitations on isolated domains of processing or unintended stimulus complexity, but rather arise from representational conflict between competing task-relevant dimensions.

4.2 Model Performance In Relation to Scaling

In human psychophysics, reaction time (RT) is often treated as a proxy for the amount of cognitive or computational resources deployed on a given task. However, the commonly used free-response paradigm—in which participants complete each trial at their own discretion—has long been recognized as a problematic index of underlying processing demands [55, 56]. Free RT is confounded by inter-individual differences in speed–accuracy trade-offs, general processing speed, and response strategies. Fast but error-prone participants may appear more efficient than slower but more accurate ones, and response latency often reflects motoric hesitation or habitual timing rather than the true time required for task-relevant computations [57, 58]. As such, free RT does not reliably reflect the processing time (PT)—the actual interval during which cognitive operations are being executed—which is the more direct indicator of computational effort. Against this backdrop, recent work has reintroduced the forced-response (paradigm into psychophysics research as a more principled alternative for estimating PT [59, 60]. In this method, response time is held constant across trials by requiring participants to respond exactly at a predetermined moment (e.g., in synchrony with a fixed "go" cue). PT is manipulated by varying the stimulus onset relative to this response cue, and the dependent measure is response accuracy as a function of available PT [61]. By decoupling the decision of when to respond from the process of deciding what to respond, the FR paradigm provides a clearer window into human performance in conflict tasks underlying limited computational resources.

In this study, we adopt the traditional free-response paradigm, as current evaluation frameworks for VLMs do not support the systematic manipulation of processing time (PT) as required by the forced-response paradigm. However, an alternative—and arguably more direct—proxy for estimating the computational resources allocated to a task is the number of parameters in a model. A central principle in machine learning holds that scaling up model size, typically measured in parameters, yields systematic improvements in reasoning and generalization capabilities [35, 62]. To explore this relationship, here we plotted model accuracy across conditions in both the Standard and Squared task sets as a function of the logarithm of model parameter size (in billions). This yielded a fine-grained mapping of VLM performance under cognitive interference relative to computational resource allocation. While we do not manipulate PT within a single model, aggregating across models of varying scale allows us to approximate a cross-sectional processing curve—analogous in form to the

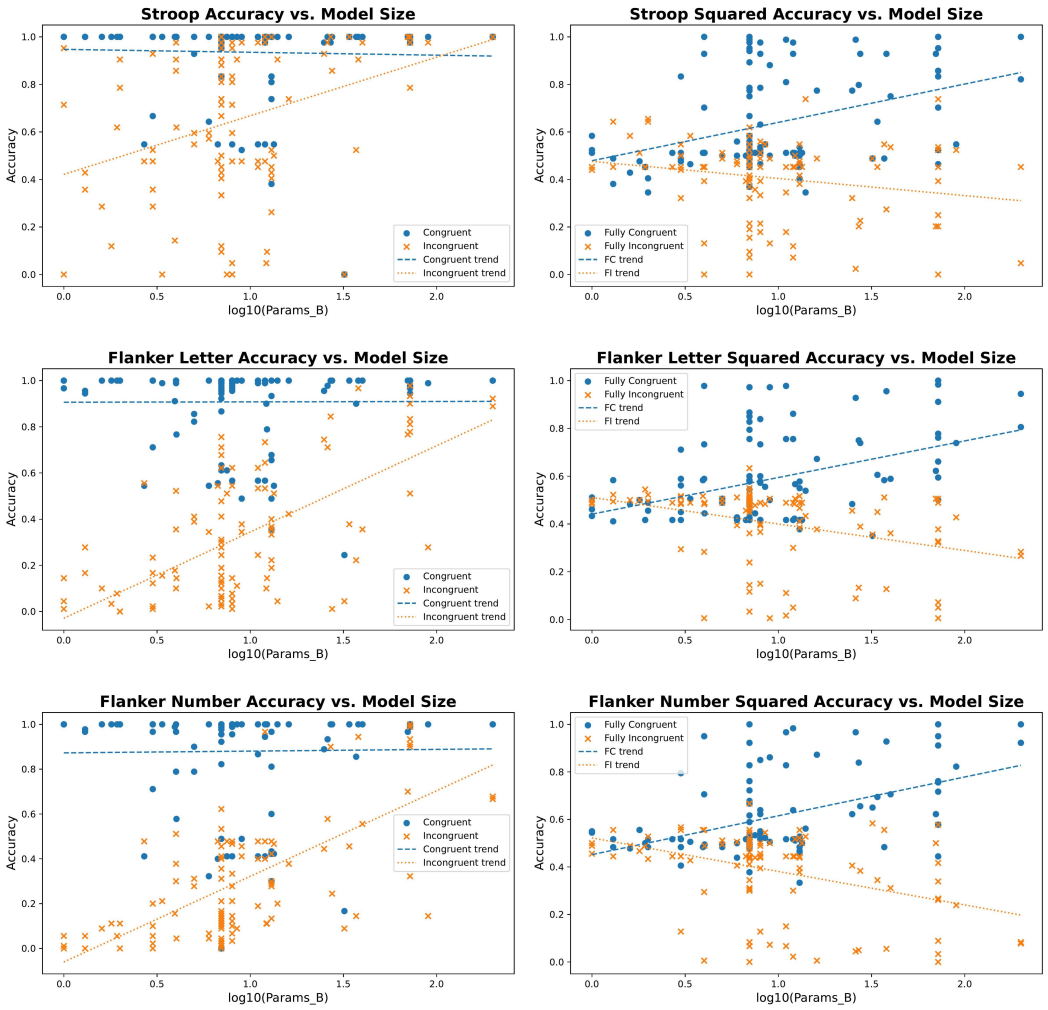

Figure 5: **Model Performance in Relation to Scaling**

time-course functions observed in human forced-response paradigms. We did not include SCRI and SIRC conditions in the scaling analysis because there is currently a lack of targeted analysis regarding performance on these conditions in relation to processing time in the human psychophysics literature.

Our results showed a striking distinction of performance patterns across model scales. On Standard tasks, the majority of models (around 0.5 to 1.5 in log parameters, by billion) revealed a robust congruency effect, with near-ceiling performance on C trials and below-chance performance on IC trials. At the higher end, the largest models (around 2.0) were the top performers, achieving ceiling-level performance on both trial types. This pattern is a direct analog to human performance under relatively high PT, ranging from high susceptibility to conflict (while still having enough resources to complete standard processing in non-conflicted scenarios) to full conflict resolution. Deviating from human performance, the smallest models (around 0 to 0.5) performed near floor on IC trials while retaining near-ceiling performance on C trials. However, this is expected, as noted in Section 3.1.2, where we suggested that a model with a strong bias toward color recognition may perform well in the standard Stroop task but fail to resolve the layered conflicts in its squared counterpart. Unlike human conscious processing—which takes approximately 300–350 ms to perform baseline functions such as color perception or character recognition—models with a strong bias toward a particular processing domain may directly prioritize said domain and completed the tasks under very limited computational resources, thereby performing near-perfect on C trials but completely subjected to conflict in I trials.

The results for Squared tasks stand in direct contrast to the patterns observed in the Standard tasks, further revealing robust individual differences between models under increased cognitive demand—precisely the motivation for introducing these tasks. Specifically, large models exhibited a pronounced congruency effect, with relatively high accuracy on FC trials and below-chance performance on FI trials. This pattern corresponds to the performance of mid-sized models on the Standard tasks. Medium-scale models mostly performed at chance, with a few approaching the behavior of larger models—exhibiting high sensitivity to conflict but incomplete resolution. The smallest models now performed strictly at chance across conditions, indicating a complete inability to execute the necessary reasoning steps. This outcome is expected, given that the Squared tasks require integrating and selectively accessing multiple sources of information within a trial—demands that cannot be met by models lacking sufficient computational resources. Notably, no model achieved near-ceiling accuracy on both FC and FI trials, indicating that the Squared tasks successfully introduced a higher level of control demand and representational conflict. Whereas model performance on Standard tasks approximates human behavior under relatively high PT, performance on Squared tasks aligns more closely with human behavior under lower PT, where increased cognitive demand requires greater computational resources—effectively shifting the functional mappings forward. Notably, model performance on the two types of Flanker tasks in relation to scaling exhibit highly corresponding performance patterns, again indicating strong convergent validity.

5 Discussion

5.1 Key Findings

Cognitive control is a core component of intelligent behavior, supporting flexible, goal-directed processing under conditions of conflict and limited resources. In this paper, we presented the first systematic evaluation of cognitive control in VLMs using a structured battery of conflict tasks. Our results show that VLMs exhibit human-like congruency effects. More complex squared tasks further revealed robust individual differences, particularly in models' sensitivity to hierarchical interference. Comparing models with different parameter counts across tasks of varying conflict levels uncovered performance patterns that closely resemble human cognitive control behavior under varying computational constraints, as indexed by processing time. This suggests a cross-system representational alignment, pointing to the emergence of cognitive control mechanisms in large-scale neural systems.

Since the early days of artificial neural network research, it has been argued that such systems lack the capacity to capture the systematicity and compositionality that define human cognition [63]. This critique was acknowledged by the symbolic cognitive architecture tradition, which sought to model intelligence through explicit representations, working memory buffers, and rule-based control mechanisms. Within this framework, cognitive control has been viewed as essential for enabling humans to apply abstract rules, manage competing goals, and adapt flexibly to novel situations. Our study provides direct empirical evidence relevant to this longstanding debate by showing that modern VLMs—trained through large-scale associative learning—can exhibit fine-grained behavioral signatures of cognitive control. This finding has both theoretical and practical significance, as it not only provides a preliminary proof of concept for the view that cognitive control—defined by flexibility, context-sensitivity, and goal-directed adaptability—can emerge from general-purpose associative learning [19], but also informs the prospect of contemporary foundation models as a paradigm for artificial general intelligence.

5.2 Limitations and Future Directions

Future work could extend this framework to include a broader range of conflict types, such as stimulus–response and action-based conflict, as the current tasks primarily target stimulus–stimulus interference. Expanding the scope of task designs would allow for a more comprehensive evaluation of control mechanisms under conditions that more closely resemble real-world demands. Additionally, further investigation is needed to determine how performance on cognitive control tasks relates to models' effectiveness in applied reasoning and decision-making settings, where successful behavior often depends on managing conflicting goals, shifting contexts, and dynamic inputs.

6 Conclusion

We showed that model performance closely aligns with human behavior under resource constraints and reveals robust individual differences across task conditions and model scales. These results provide substantial support for the emergence of human-like cognitive control in current multi-modal foundation models.

Acknowledgments and Disclosure of Funding

We thank Han Zhang, Jacob Sellers, and Sarah Liberatore for their insights and comments. We thank Zillion Network Inc. for providing the computation used in this work. Their optimized peak/off-peak scheduling, high-throughput storage infrastructure, and automated environment management enabled the cost-efficient and reliable execution of our experiments.

References

[1] Jens Rasmussen. The role of error in organizing behaviour. *Ergonomics*, 33(10-11):1185–1199, 1990.

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

[4] David Badre. Cognitive control. *Annual Review of Psychology*, 76, 2024.

[5] John R Anderson. *The architecture of cognition*. Psychology Press, 1983.

[6] Jacob Russin, Randall C O'Reilly, and Yoshua Bengio. Deep learning needs a prefrontal cortex. *"Bridging AI and Cognitive Science" Workshop at the International Conference on Learning Representation (ICLR)*, 107(603-616):1, 2020.

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

[8] Adele Diamond. Executive functions. *Annual review of psychology*, 64(1):135–168, 2013.

[9] Kenneth R Hammond and David A Summers. Cognitive control. *Psychological review*, 79(1): 58, 1972.

[10] Todd S Braver. The variable nature of cognitive control: a dual mechanisms framework. *Trends in cognitive sciences*, 16(2):106–113, 2012.

[11] Kenji Matsumoto and Keiji Tanaka. Conflict and cognitive control. *Science*, 303(5660):969–970, 2004.

[12] Wim Notebaert, Wim Gevers, Frederick Verbruggen, and Baptist Liefooghe. Top-down and bottom-up sequential modulations of congruency effects. *Psychonomic bulletin & review*, 13 (1):112–117, 2006.

[13] Frederick Verbruggen, Wim Notebaert, Baptist Liefooghe, and André Vandierendonck. Stimulus- and response-conflict-induced cognitive control in the flanker task. *Psychonomic Bulletin & Review*, 13:328–333, 2006.

[14] Steven J Luck and Edward K Vogel. The capacity of visual working memory for features and conjunctions. *Nature*, 390(6657):279–281, 1997.

[15] Edward Awh, Brian Barton, and Edward K Vogel. Visual working memory represents a fixed number of items regardless of complexity. *Psychological science*, 18(7):622–628, 2007.

[16] Edward E Smith and John Jonides. Storage and executive processes in the frontal lobes. *Science*, 283(5408):1657–1661, 1999.

[17] Randall W Engle. Working memory capacity as executive attention. *Current directions in psychological science*, 11(1):19–23, 2002.

[18] Jonathan D Cohen, Kevin Dunbar, and James L McClelland. On the control of automatic processes: a parallel distributed processing account of the stroop effect. *Psychological review*, 97(3):332, 1990.

[19] Elger Abrahamse, Senne Braem, Wim Notebaert, and Tom Verguts. Grounding cognitive control in associative learning. *Psychological bulletin*, 142(7):693, 2016.

[20] Tom Verguts and Wim Notebaert. Adaptation by binding: A learning account of cognitive control. *Trends in cognitive sciences*, 13(6):252–257, 2009.

[21] Earl K Miller and Jonathan D Cohen. An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, 24(1):167–202, 2001.

[22] John Duncan. The structure of cognition: attentional episodes in mind and brain. *Neuron*, 80 (1):35–50, 2013.

[23] Etienne Koechlin, Chrystele Ody, and Frédérique Kouneiher. The architecture of cognitive control in the human prefrontal cortex. *Science*, 302(5648):1181–1185, 2003.

[24] Dario D Salvucci and Niels A Taatgen. Toward a unified view of cognitive control. *Topics in cognitive science*, 3(2):227–230, 2011.

[25] Allen Newell, Paul S Rosenbloom, and John E Laird. Symbolic architectures for cognition. 1989.

[26] John R Anderson. *The architecture of cognition*. Psychology Press, 1982.

[27] Paul Thagard. Cognitive architectures. *The Cambridge handbook of cognitive science*, 3:50–70, 2012.

[28] Pat Langley, John E Laird, and Seth Rogers. Cognitive architectures: Research issues and challenges. *Cognitive Systems Research*, 10(2):141–160, 2009.

[29] Iuliia Kotseruba and John K Tsotsos. 40 years of cognitive architectures: core cognitive abilities and practical applications. *Artificial Intelligence Review*, 53(1):17–94, 2020.

[30] Richard M Young and Richard L Lewis. The soar cognitive architecture and human working memory. *Models of working memory: Mechanisms of active maintenance and executive control*, pages 224–256, 1999.

[31] John E Laird. *The Soar cognitive architecture*. MIT press, 2019.

[32] John R Anderson, Michael Matessa, and Christian Lebiere. Act-r: A theory of higher level cognition and its relation to visual attention. *Human–Computer Interaction*, 12(4):439–462, 1997.

[33] Frank E Ritter, Farnaz Tehranchi, and Jacob D Oury. Act-r: A cognitive architecture for modeling cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 10(3):e1488, 2019.

[34] Bryan Stearns and John E Laird. Toward unifying cognitive architecture and neural task set theories. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 42, 2020.

[35] Richard Sutton. The bitter lesson. *Incomplete Ideas (blog)*, 13(1):38, 2019.

[36] Earl K Miller. The prefontral cortex and cognitive control. *Nature reviews neuroscience*, 1(1): 59–65, 2000.

[37] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308, 2024.

[38] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[39] Gemini. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv: 2312.11805*, 2023.

[40] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv: 2306.13394*, 2023.

[41] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv: 2303.08774*, 2023.

[42] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024.

[43] Wenxiao Cai, Yaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. *arXiv preprint arXiv:2406.13642*, 2024.

[44] Yuliang Liu, Zhang Li, Biao Yang, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023.

[45] Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-llm: Extending language model for 3d visual understanding and reasoning. *arXiv preprint arXiv:2403.11401*, 2024.

[46] Yijiang Li, Qingying Gao, Tianwei Zhao, Bingyang Wang, Haiyun Lyu, Luo Dezhi Sun, Haoran, and Hokin Deng. Core knowledge deficits in multi-modal language models. *arXiv preprint arXiv:2410.10855*, 2025.

[47] Penghao Wu and Saining Xie. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094, 2024.

[48] Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024.

[49] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. *arXiv preprint arXiv:2403.16999*, 2024.

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

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

[52] Barbara A Eriksen and Charles W Eriksen. Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & psychophysics*, 16(1):143–149, 1974.

[53] Alexander P Burgoyne, Jason S Tsukahara, Cody A Mashburn, Richard Pak, and Randall W Engle. Nature and measurement of attention control. *Journal of Experimental Psychology: General*, 152(8):2369, 2023.

[54] Craig Hedge, Georgina Powell, and Petroc Sumner. The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior research methods*, 50: 1166–1186, 2018.

[55] Robert G Pachella. The interpretation of reaction time in information-processing research 1. In *Human information processing*, pages 41–82. Routledge, 2021.

[56] Christopher Draheim, Cody A Mashburn, Jessie D Martin, and Randall W Engle. Reaction time in differential and developmental research: A review and commentary on the problems and alternatives. *Psychological bulletin*, 145(5):508, 2019.

[57] Adrian M Haith, Jina Pakpoor, and John W Krakauer. Independence of movement preparation and movement initiation. *Journal of Neuroscience*, 36(10):3007–3015, 2016.

[58] Aaron L Wong, Jeff Goldsmith, Alexander D Forrence, Adrian M Haith, and John W Krakauer. Reaction times can reflect habits rather than computations. *Elife*, 6:e28075, 2017.

[59] Barbara Anne Dosher. The retrieval of sentences from memory: A speed-accuracy study. *Cognitive psychology*, 8(3):291–310, 1976.

[60] Wayne A Wickelgren. Speed-accuracy tradeoff and information processing dynamics. *Acta psychologica*, 41(1):67–85, 1977.

[61] Taraz G Lee, Jacob Sellers, John Jonides, and Han Zhang. The forced-response method: A new chronometric approach to measure conflict processing. *Behavior Research Methods*, 57(1): 1–14, 2025.

[62] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[63] Jerry A Fodor and Zenon W Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988.