Towards Contextual Robot Intent Explanation for Hierarchical Vision–Language–Action Collaboration

Amanuel Ergogo, Thuc Anh Nguyen, Mark Anjoul, Aarav Jain, Yasa Zaheen, Divyamshu Shrestha, Zhao Han

***Abstract*— Assistive and collaborative robots must coordinate their actions with human partners, make their intent interpretable, and accept mid-task interruptions in shared environments with shared roles, such as collaborative medication dispensing and collaborative cooking. Current vision-language-action (VLA) models achieve high task success in single-agent settings but lack mechanisms to communicate intent, express confidence, or adapt when collaborating with humans, as VLA models' plans reside in a latent space that no external agent can inspect. In this work in progress, we propose CRIE (Contextual Robot Intent Explanation), a hierarchical planner designed to coordinate tasks across human and robot agents and generate intent utterances grounded in VLA action embeddings, execution progress, and entropy-based uncertainty. We also outline a planned evaluation to test CRIE's improvement.**

I. INTRODUCTION

Assistive and collaborative robots are increasingly needed in settings where humans and machines share tasks, environments, and responsibility: supporting with daily assistive tasks, coordinating with clinical staff during medication preparation so that they focus on patient-centered tasks, or assisting individuals with motor impairments, for example, collaborative cooking to create a sense of independence [1]. For such collaboration to succeed, robots must legibly coordinate their actions with a human partner, make their intent and confidence interpretable, and accept mid-task interruptions with dynamic replanning [2], [3].

However, current learned manipulation policies do not meet these demands. These policies achieve high task success in controlled, single-agent settings, but coordination degrades sharply during collaboration. In our prior 32-participant, 288-trial collaborative medication-dispensing study (under review), we found that task success dropped from 99% without a human partner to 81% and 75% for ACT and Diffusion policy, respectively. Grounded video coding identified 127 coordination failures across eight categories, i.e., Passive Wait, Redundant Retrieval, Missed Grab, Slippage, Task Model Uncertainty, Capability Miscalibration, Safety Avoidance, and Safety Conflict. We aim to solve dominant failure cases by clustering into three design needs. (1) Passive Wait (27.6%) and Redundant Retrieval (16.5%) together account for 44.1% of all events, calling for **intent legibility**. (2) Task Model Uncertainty (10.2%) and Capability Miscalibration (10.2%) together account for 20.4%, indicating a need for **capability and uncertainty calibration**. (3) Safety Avoidance (8.7%) and Safety Conflict (5.5%) account for 14.2%, motivating **interruptible replanning**.

All authors are with the Bellini College of Artificial Intelligence, Cybersecurity and Computing, University of South Florida, Tampa, FL 33620, USA. {aergogo, zhaohan}@usf.edu

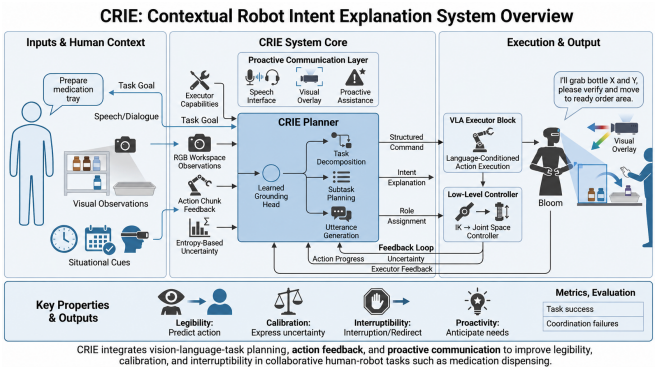

Fig. 1. CRIE system overview. The CRIE Planner receives task goals, workspace observations, action chunk feedback, and entropy-based uncertainty, and produces structured commands, intent explanations, and role assignments via a learned grounding head. A proactive communication layer delivers intent through speech and visual overlay. (AI-assisted figure, reviewed by authors.)

In this work in progress, we propose **CRIE** (Contextual Robot Intent Explanation), a hierarchical architecture centered on the CRIE Planner, designed to bridge capable manipulation and effective human-robot coordination (Fig. 1). The CRIE Planner is intended to decompose tasks and plan subtask execution across human and robot agents, and generate intent utterances grounded in VLA executor capabilities, workspace observations, action chunk feedback, and uncertainty. A learned grounding head is designed to ground planner reasoning directly to VLA's internal representations, enabling detection and communication of executor-side failures. A proactive communication layer is designed to complement the planner by sustaining natural conversation, monitoring situational cues to anticipate shared task goals, triggering the task coordinator, and delivering intent through speech and visual overlay [4]. We also outline a planned evaluation within a coordination-failure taxonomy.

Our main contributions are:

1) **CRIE Planner:** A planner designed to decompose collaborative tasks across human and robot agents and generate intent utterances grounded in VLA action embeddings, execution progress, and entropy via a learned grounding head.
2) **Proactive Communication Layer:** A system component intended to sustain social interaction, monitor situational context to detect task intent from dialogue and situational context, then trigger the CRIE Planner

for physical assistance, and deliver robot intent through speech and projected visual overlay.

3) **Evaluation Framework:** A planned two-level evaluation comparing three planner variants across task decomposition, subtask planning, and utterance quality in human-agent coordination, and assessing the full system in a pilot study using an 8-category coordination failure taxonomy.

II. Related Work

VLA models. ACT [5] uses conditional VAEs over action chunks for fine-grained bimanual manipulation, Diffusion Policy [6] casts action generation as iterative denoising, and more recent vision-language-action methods add language conditioning for cross-task and cross-embodiment generalization [7]–[9]. Despite these advances, such systems are generally evaluated as single-agent executors and provide no explicit mechanism for exposing intent, confidence, or other coordination-relevant execution signals to a human collaborator. CRIE addresses this gap by introducing a human-agent coordinator and communication layer that operates alongside the executor to generate grounded, intent explanations.

Hierarchical language-conditioned control. Hierarchical language-conditioned control decompose high-level natural-language instructions into a hierarchy of sub-goals or skill primitives: a high-level planner reasons to select what to do, while a low-level policy executes how to do it in continuous action space [10]. SayCan [8] grounds LLM proposals in learned affordances; Code as Policies [9] compiles LLM output to executable programs; Inner Monologue [11] feeds environment feedback to an LLM for replanning. These systems demonstrate the value of language as a structuring medium for robotic control but treat it exclusively as an internal command channel. CRIE extends this paradigm by using hierarchical decomposition as the basis for outward-facing, human-directed explanations using language as a *coordination medium* between agents.

Transparency and explainability in HRC. Dragan et al. [3] showed that motion trajectories can be optimized for legibility, enabling observers to infer goals from early motion cues effective for short reaches but not multi-step tasks requiring a full subtask plan. Han et al. [12] generated natural-language explanations from behavior trees; CRIE generates explanations from learned policies via the VLM intermediary, requiring no hand-authored task model.

III. CRIE System Design

CRIE is designed around three modules (Fig. 2): (1) the CRIE Planner, which decomposes the task goal into subtasks, allocates them across the human–robot team, and generates intent utterances to improve legibility; to support calibration, these utterances are grounded in VLA capabilities, execution progress, and entropy-based uncertainty, enabling the robot to communicate confidence and request assistance when needed; (2) a VLA Executor with low-level control for continuous action generation; and (3) a Proactive Communication Layer for speech, visual intent display, and feedback-driven replanning that captures mid-task verbal redirects, thereby improving interruptibility.

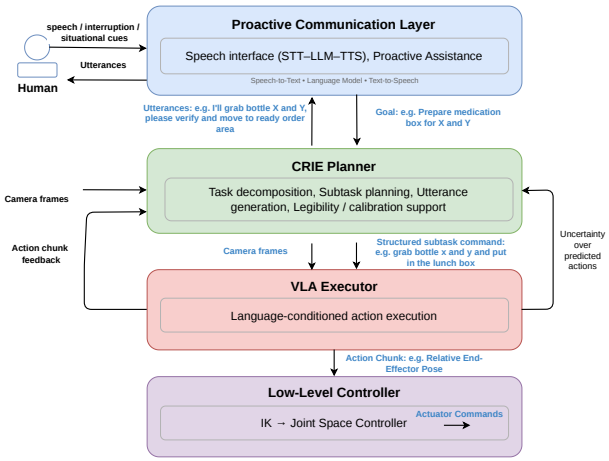

Fig. 2. CRIE architecture and data flow. Structured subtask commands flow from the planner to the VLA executor and low-level controller. Executor feedback, including progress and entropy, is returned to the planner for replanning, calibration, and interruption handling. The communication layer delivers planner-generated utterances through speech and visual overlay.

A. CRIE Planner

The CRIE Planner serves as the central reasoning component. It is a VLM-based module designed to map multimodal context to both a structured subtask command and a human-directed explanation. The command takes the form: $c_t = \langle verb,\ object,\ source,\ target,\ constraint \rangle$. This flat, parseable format serves a dual purpose: It constrains the VLA executor and provides the semantic content from which human-facing utterances are generated without requiring a second model call. For example, (pick, ibuprofen, shelf_2, tray, verify_label) produces the utterance "I'll pick the ibuprofen from shelf 2. Please verify the label."

The planner takes five input streams that together ground its reasoning in task context and executor's internal state:

1) **Executor capabilities:** A structured description of what manipulation skills the VLA can perform, enabling the planner to assign subtasks appropriately across the human–robot team.
2) **Task goal:** The high-level objective (e.g., "prepare morning medication tray"), provided by the user or generated proactively by the communication layer.
3) **RGB workspace observations:** Camera capturing environment changes caused by actions, enabling the planner to track progress and detect unexpected states.
4) **Action chunk feedback:** The VLA executor's generated action chunks, projected into the VLM's input space via a learned grounding head to reason what the executor is *doing*, not just what it was *told* to do.
5) **Entropy-based uncertainty:** Action-distribution entropy estimates executor uncertainty. We calibrate τ_h on held-out successful policy rollouts as the 95th percentile of entropy across action chunks [13]. At

runtime, entropy above τ_h triggers an uncertainty utterance (e.g., "I'm not sure about this bottle; can you confirm?") and trigger replanning.

From these inputs, the planner is designed to produce three outputs per subtask transition: (i) the structured command c_t directing the VLA executor, (ii) a human-directed utterance for legibility and calibration (e.g., explaining the current subtask, expressing uncertainty, or requesting human action), and (iii) a role assignment indicating whether the subtask is robot-executed, human-executed, or jointly performed.

1) Planners: We plan to compare three conditions to measure how grounding depth affects coordination quality:

V1 – Zero-shot. In-context prompting with no fine-tuning, establishing a baseline for off-the-shelf VLM planning.

V2 – Fine-tuned. Fine-tuned on annotated (observation, command) pairs from expert demonstrations, with structured-output constraints to ensure valid command format.

V3 – Action-grounded. V2 augmented with the three executor feedback streams (action chunk feedback, execution progress, entropy) which are tokenized by a grounding head trained on a mixture of synthetic and human-labeled data and concatenated to the planner input. This variant is designed to detect executor-side failures that are invisible to vision-only planners and to generate uncertainty-calibrated utterances grounded in the executor's actual confidence.

B. VLA Executor and Low-Level Control

The executor receives the structured command c_t and produces continuous action chunks (relative end-effector poses). A low-level IK joint-space controller converts end-effector poses into actuator commands. Output actions and uncertainty feedback are returned to the CRIE Planner and ingested for replanning and uncertainty communication.

C. Proactive Communication Layer

The communication layer manages social interaction and delivers planner-generated utterances with two functions:

Intent delivery. Planner-generated utterances are communicated through two channels: speech (via STT-LLM-TTS pipeline) and visual overlay (bounding box and approach-direction arrow projected onto the workspace via an integrated projector [4]). An adaptive gate suppresses output when the subtask and confidence are unchanged, firing on subtask transitions, confidence drops ($h_t > \tau_h$), or sustained partner inactivity. This reduces message frequency while preserving coverage of coordination-relevant events.

Proactive goal anticipation. The proactive goal-anticipation layer uses conversation and situational cues, including time of day, user routines, and visual context, to infer task goals and offer assistance. When a goal is detected, it sends a capability-grounded textual goal to the CRIE Planner. During execution, it handles verbal interrupts by updating the planner with redirected goals and triggering replanning. This proactivity reduces the coordination burden placed on users by passive command-driven robots [14].

1) Task Detection: The robot must distinguish task-directed requests from casual dialogue. We embed detection within the conversational LLM: the system prompt instructs the model to prepend a structured tag upon detecting task intent (`[TASK] ⟨goal⟩ | ⟨response⟩`). A preliminary proof-of-concept test on a 27-utterance testbed showed 89% accuracy (F1: 0.91) versus 52% (F1: 0.43) for a keyword baseline, with perfect recall ensuring no missed task requests. Full validation is planned as part of the system evaluation.

2) Interrupt Handling: Mid-task verbal redirects (e.g., "Stop, get the blue cup instead") are expected to be captured by speech recognition. The transcript is added to the planner's dialogue history, triggering a new command. The layer acknowledges the change (e.g., "switching to the blue cup"). Each redirect is logged as a preference pair $(c_{\text{original}}, c_{\text{redirect}} \mid s_t)$ for future direct preference optimization fine-tuning [15].

D. Bloom Robot Platform

We will deploy CRIE on a modified bimanual Bloom social robot designed for expressive interaction and physical task execution. The platform features bimanual 6-DoF arms; a head with 6-DoF orientation control and customizable 3D-printed face; an integrated tiny projector for visual overlays, ReSpeaker microphone array for beamforming, and 5W stereo speakers [16]; a Raspberry Pi 5 with AI Hat for local real-time speech processing (proactive communication layer inference); and battery-powered UPS for untethered operation in home and clinical environments while the CRIE planner and vision-language reasoning run on a desktop.

IV. HYPOTHESES

As CRIE will address the three coordination needs: intent legibility, capability, and uncertainty calibration, and interruption-aware replanning. We believe that increasing planner grounding and adding explicit intent communication will improve both planner-level quality (H1) and human-robot collaboration outcomes (H2–H6).

H1–Improved Planner Quality. The action-grounded planner (V3) will outperform the fine-tuned planner (V2), which will outperform the zero-shot planner (V1), on task decomposition accuracy, subtask plan validity, and explanation quality, defined by relevance, accuracy, and informativeness.

H2–Improved Legibility. Relative to the zero-shot planner (V1), CRIE will improve robot behavior legibility, yielding fewer *Passive Wait count* and *Redundant Retrieval count*.

H3–Improved Calibration. Relative to V1 and V2, humans will be better able to calibrate the action-grounded planner (V3)'s capabilities, yielding fewer *Task Model Uncertainty* and *Capability Miscalibration* errors.

H4–Improved Interruptibility. Relative to V1 and V2, CRIE communication will achieve a higher *redirect success rate* when interrupted.

H5–Improved Collaboration Performance. CRIE will improve task performance.

H6–Improved Overall Collaboration Fluency. CRIE produces higher collaborative fluency than V1 and V2.

Additionally, as VLA may produce inaccurate information like LLM, we will conduct additional analysis on hallucination rate, the fraction of robot utterances with false claims relative to the planner command, executor state, or workspace observation [17].

V. PLANNED EVALUATION

To test the hypotheses, we plan a two-level evaluation on the CRIE Planner in isolation and the fully integrated system.

A. Conditions

At the *planner level*, we compare the three variants (V1: zero-shot, V2: fine-tuned, V3: action-grounded) on held-out scenarios to isolate the effect of grounding depth on planning quality. At the *system level*, we cross the planner variant with the communication condition (speech-only, visual-overlay-only, both) in a within-subjects 30-participant study to measure how each combination affects coordination.

B. Tasks

Collaborative medication dispensing. Human-robot agents fulfill an order for two medications in a shared workspace with a bottle shelf, labeling station, and Ready Order Area. Because medication dispensing is safety-critical and requires tightly coupled coordination, the robot's future actions are often difficult for the human partner to anticipate from behavior alone. The task, therefore, highlights the need for explicit intent communication, uncertainty-aware coordination, and online replanning in shared workspaces.

Collaborative cooking. We will also evaluate CRIE on a collaborative cooking task adapted from Overcooked-AI [18]. This task captures human–agent collaboration dynamics in shared domestic and assistive settings, requiring coordination of multi-step food-preparation actions such as fetching ingredients, placing items in shared work areas, sequencing preparation steps, and timing handoffs. It requires humans and agents to switch roles, handle handoffs, and make online coordination decisions.

C. Metrics

We will measure **planner quality** by *task decomposition accuracy*, the fraction of generated subtask sequences that match the ground-truth decomposition in step count and ordering, *Subtask plan validity*: the fraction of commands with correct role assignment. *Utterance quality* three 5-point Likert dimensions: relevance, accuracy, and informativeness.

We will measure **legibility** by Passive Wait count, number of events both agents waited to confirm their future actions and Redundant Retrieval count, the number of times the agents duplicate an action, both identified via video coding. *Calibration* (Task Model Uncertainty, Capability Miscalibration): unnecessary takeover rate ($N_{\text{takeovers}}/N_{\text{subtasks}}$), capturing under-trust; and missed intervention rate ($N_{\text{unaddressed}}/N_{\text{failures}}$).

We will measure **interruptibility** (Safety Avoidance, Safety Conflict) by redirect success rate ($N_{\text{success}}/N_{\text{probes}}$) on scripted verbal redirect.

We will measure **task performance** by completed deliveries per episode, completion time, and safety violation count.

Fluency will be measured by the eight-item widely-used scale consisting of objective and subjective metrics by Hoffman [19].

Free-form responses will be collected to assess how design choices, including VLM planning, VLA execution, action grounding, speech, visual overlays, proactive goal detection, and interrupt handling, affects system performance.

D. Data Analysis

We will test H1-H5 using Bayesian repeated-measures ANOVA [20], which quantifies evidence for and against an effect via the Bayes Factor (BF). We interpret BF_{10} using the classification of Wagenmakers et al. [20]: anecdotal $(1, 3]$, moderate $(3, 10]$, strong $(10, 30]$, very strong $(30, 100]$, and extreme $(100, \infty)$; inverted intervals apply for evidence favoring H_0. When a main effect is supported ($BF_{10} > 3$) or cannot be ruled out ($BF_{10} \in [\frac{1}{3}, 3]$), we conduct post hoc pairwise comparisons. Planner-level metrics are tested on held-out scenarios; system-level metrics are tested on participant data. We will thematically code free-form responses to identify how action grounding, speech, visual overlays, proactive goal detection, and interrupt handling shaped coordination outcomes.

VI. DISCUSSION

CRIE aims to target three coordination needs identified in our prior failure analysis: intent legibility, capability and uncertainty calibration, and interruption-aware replanning by introducing a planner and communication layer that translate latent policy state into partner-facing language. The structured command $\langle \textit{verb}, \textit{object}, \textit{source}, \textit{target}, \textit{constraint} \rangle$ both constrains the executor and supplies the content of the human-directed utterance, keeping action and explanation consistent. CRIE differs from prior hierarchical language-conditioned systems by treating language as an outward-facing coordination medium rather than only as an internal command channel. Also, Grounding planner reasoning in action embeddings, progress, and entropy is intended to surface executor-side uncertainty that open-loop execution cannot detect, while role assignment and the gated speech–overlay channel together address Passive Wait, Redundant Retrieval, and mid-task redirects. Limitations remain: the system is not yet evaluated, and entropy might be an unreliable proxy for uncertainty that requires model-specific calibration.

VII. CONCLUSION

We proposed CRIE, a hierarchical planner and proactive communication architecture that grounds structured commands and intent utterances in task context, executor feedback, and uncertainty, treating language as both an internal command interface and an outward-facing coordination medium. A planned two-level evaluation will compare planner variants and communication modalities on collaborative medication dispensing and cooking, using our eight-category failure taxonomy to assess whether grounded explanation reduces coordination failures and improves fluency.

REFERENCES

[1] A. Ergogo, D. Dall'Alba, and P. Korzeniowski, "Veragmil: Virtual environment for scooping granular foods with imitation learning models," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025, pp. 13 075–13 082.

[2] N. Sebanz, H. Bekkering, and G. Knoblich, "Joint action: Bodies and minds moving together," *Trends in Cognitive Sciences*, vol. 10, no. 2, pp. 70–76, 2006.

[3] A. D. Dragan, K. C. T. Lee, and S. S. Srinivasa, "Legibility and predictability of robot motion," in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2013, pp. 301–308.

[4] Z. Han, "Anywhere projected AR for robot communication," in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2025.

[5] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," in *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[6] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," in *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[7] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi *et al.*, "OpenVLA: An open-source vision-language-action model," in *Proceedings of the Conference on Robot Learning (CoRL)*, 2024.

[8] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gober, K. Hausman *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," in *Proceedings of the Conference on Robot Learning (CoRL)*, 2022.

[9] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, "Code as policies: Language model programs for embodied control," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[10] J. Sun, A. Curtis, Y. You, Y. Xu, M. Koehle, Q. Chen, S. Huang, L. Guibas, S. Chitta, M. Schwager *et al.*, "Arch: Hierarchical hybrid learning for long-horizon contact-rich robotic assembly," *arXiv preprint arXiv:2409.16451*, 2024.

[11] W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar *et al.*, "Inner monologue: Embodied reasoning through planning with language models," in *Proceedings of the Conference on Robot Learning (CoRL)*, 2023.

[12] Z. Han, D. Giger, J. Allspaw, M. S. Lee, H. Admoni, and H. A. Yanco, "Building the foundation of robot explanation generation using behavior trees," in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2021.

[13] C. Xu, T. K. Nguyen, E. Dixon, C. Rodriguez, P. Miller, R. Lee, P. Shah, R. Ambrus, H. Nishimura, and M. Itkina, "Can we detect failures without failure data? uncertainty-aware runtime failure detection for imitation learning policies," *arXiv preprint arXiv:2503.08558*, 2025.

[14] M. Ali, S. Alili, M. Warnier, and R. Alami, "An architecture supporting proactive robot companion behavior," *SSAISB: The Society for the Study of Artificial Intelligence and the Simulation of Behaviour. 0 (0)*, pp. p1–8, 2009.

[15] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[16] Z. Han, "Bloom preview: A low-cost LLM-powered social robot," in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2026.

[17] C. Y. Kim, C. P. Lee, and B. Mutlu, "Understanding large-language model (llm)-powered human-robot interaction," in *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 371–380. [Online]. Available: https://doi.org/10.1145/3610977.3634966

[18] M. Carroll, R. Shah, M. K. Ho, T. L. Griffiths, S. A. Seshia, P. Abbeel, and A. Dragan, "On the utility of learning about humans for human-AI coordination," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[19] G. Hoffman, "Evaluating fluency in human–robot collaboration," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 3, pp. 209–218, 2019.

[20] E.-J. Wagenmakers, M. Marsman, T. Jamil, A. Ly, J. Verhagen, J. Love, R. Selker, Q. F. Gronau, M. Šmíra, S. Epskamp *et al.*, "Bayesian inference for psychology. part i: Theoretical advantages and practical ramifications," *Psychonomic bulletin & review*, vol. 25, no. 1, pp. 35–57, 2018.