

VLI

Vision-Language-Interaction: Building a System to Interact, Not to Obey

A research position on making interaction, rather than task-completion, the objective of embodied agency — and metering every engagement with the world.

David Jean Charlot, PhD

Dean of Physical AI · The Charlot Lab, Institute for Physical AI @ BMI

Correspondence: contact@physicalai-bmi.org · physicalai-bmi.org

Interactive companion: physicalai-bmi.org/research/charlot-lab#topic-vli

ABSTRACT. Contemporary embodied models are trained to map perception and language to action, and are optimized to complete named tasks. This report argues that task-completion under-specifies what a system with agency actually needs, because a policy is only ever prepared for the situations its objective named, and the world routinely presents situations no order anticipated. It develops a research position we call Vision-Language-Interaction: make interaction itself the objective, so that a system builds, by construction, the repertoire it needs to act in situations it was never told about. We are explicit that this position is not novel in isolation; it restates, for embodied language-conditioned agents, ideas long present in the intrinsic-motivation, artificial-curiosity, and empowerment literature, and we cite that prior art rather than claim to originate it. What we add is an engineering commitment: the lab's system, Ambit, runs every interaction on the MathGround substrate, so each decision to engage the world carries a joules receipt and a replayability class. Section 2 states the review method. Sections 3 and 4 situate interaction against action and against the intrinsic-motivation tradition. Sections 5 and 6 describe the metered substrate and what a repertoire built by interaction buys. Section 7 is explicit about what is a position and what is demonstrated. The report presents no new experimental measurements.

1. Introduction

A system with agency cannot only take orders. To act in the world it has to interact with the world, and with everything in it: people, objects, surfaces, forces, signals, other systems. An order names a small, closed set of desired outcomes. The world does not. Between the moment a command is issued and the moment it would be satisfied, an embodied system

meets contact events, occlusions, slippage, interruptions, and other agents, none of which the command enumerated. The competence to handle those is not contained in the command; it has to come from somewhere else.

The prevailing design does not put it there directly. Vision-Language-Action (VLA) models map images and instructions to motor commands and are trained and evaluated on task success^[1,2]. This is a real and useful capability, and it has advanced quickly. But an order-following system is, by construction, only ever ready for the orders it was given. This report takes a different objective as its subject. A system built to maximize interaction is, by construction, assembling the repertoire it needs to act in any situation it meets, because interaction is how it meets the world and learns what it can do there. We have watched machines move with intention for a century, in the films and the anime and the toys on the shelf, long before one could be built. What is new is not the wish. It is that interaction, not command, can now be the quantity a system is built to maximize.

2. Scope and method

This is a review and a research position, not a report of experiments. It surveys the VLA literature and the older intrinsic-motivation, curiosity, and empowerment literature, and it states a position about how the two should be combined for embodied agency. For each external claim it favors a primary, verifiable source; every reference below resolves to a published paper, preprint, or the authors' technical disclosure, and the identifiers are given in the reference list. The report describes one system, Ambit, and its substrate, MathGround, at the level of design commitments; it does not report benchmark numbers, success rates, or measured energies for them, and where a capability is a prototype or a plan rather than a demonstrated result it says so plainly. The central claim of Section 4 — that interaction is the right objective — is a position, and prior art for its ingredients is cited so the novelty is not overstated. The report reports no new experimental measurements.

3. From action to interaction

The VLA formulation is clean. A model receives an image stream and a natural-language instruction and emits actions, and web-scale vision-language pretraining is transferred into control so that the instruction can range over open vocabulary^[1]. Open weights and standard recipes have made the approach reproducible and widely built upon^[2]. The objective throughout is task-completion: the policy is trained on demonstrations of tasks and scored by whether the named task succeeds.

Task-completion under-specifies readiness for a simple reason. The objective assigns value only along the trajectories the task names. Everything the system might do that is not on the path to a named goal is, to the objective, invisible; it is neither rewarded nor practiced. A policy optimized this way can become excellent at its tasks and remain empty everywhere else, so that the first unnamed situation — a dropped object, a closed door, a person who moves — falls outside the region the objective ever shaped. Recent work has started to widen the formulation from the inside. Concurrent see-hear-speak-act architectures let an agent observe,

listen, speak, and act at the same time and be interrupted mid-action, which is a move away from the single-command, single-trajectory picture toward continuous engagement^[3]. And policies are beginning to improve from their own operational experience rather than from demonstrations alone, correcting their own mistakes through repeated practice^[4]. Both are steps from action toward interaction. Vision-Language-Interaction (VLI) names the endpoint of that direction: interaction, not the completion of a named task, as the thing the system is built to maximize.

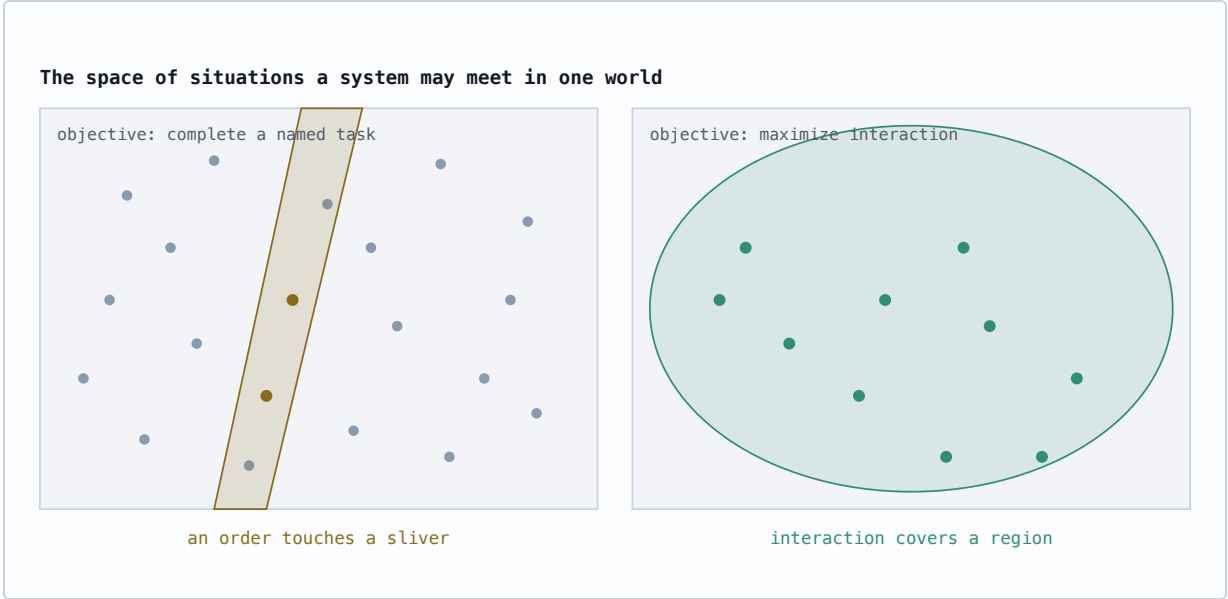


Figure 1. Two objectives over the same world. Grey dots are situations the system may meet. Left: a task objective assigns value only along the narrow band from the current state to the named goal, so most situations are never shaped by it. Right: an interaction objective spreads value across a broad region, so a repertoire accumulates before any particular situation arrives. The figure is schematic; the shapes are illustrative, not measured.

4. Interaction as the objective, honestly placed

The idea that an agent should be driven by something other than an external task is not new, and this report does not claim it is. It restates, for language-conditioned embodied models, a position with a long record in machine learning and developmental robotics. Curiosity formalized as the prediction error of an agent about the consequences of its own actions produces exploration without any task reward, and drives an agent to seek the parts of its environment it does not yet model^[5]. The broader formal theory of creativity and intrinsic motivation frames an agent as rewarded for discovering novel, learnable regularities, that is, for improving its own model of the world^[6]. A separate line makes the objective the agent's own influence over its future rather than its surprise. Empowerment measures the information-theoretic capacity of the channel from an agent's actions to its later sensory states, and proposes that an agent maximize it^[7,8]. Informally, empowerment is the number of futures an agent can reliably bring about, written

$$\mathfrak{E}(s) = \max_{p(a^n)} I(A_t^n; S_{t+n} \mid s),$$

the maximum mutual information between a sequence of actions and the resulting state. The typology of these approaches was mapped nearly two decades ago, and its authors were careful that the various proposals are not interchangeable and are often under-specified^[9]. We inherit that caution. Interaction, as an objective, is a family, not a single formula: it can be read as curiosity about what the world does when touched, as empowerment over the states the body can reach, or as the coverage in Figure 1. The position of this report is that some member of this family, not task-completion, is the right training objective for a system meant to act in an open world, because such an objective values the acquisition of a repertoire directly rather than as a byproduct. That is a claim about what to optimize, and its ingredients are prior art; the contribution here is to argue it specifically for the VLA setting and to ground it, which is the subject of the next section.

5. Grounded and metered: Ambit on MathGround

An objective that rewards engaging the world can reward engaging it wastefully. Curiosity that is free will explore forever; empowerment that is free will reach for every future indiscriminately. Interaction, if it is to be a serious objective for an embodied system, must be priced. The lab's effort toward VLI is a system named Ambit, and it is grounded in a specific way: every interaction runs on the MathGround substrate, so that each decision to engage the world carries two attached quantities. The first is a **joules receipt**: the energy attributable to the computation and, where instrumented, the actuation of that interaction, recorded rather than estimated after the fact. The second is a **replayability class**: a label stating whether the interaction, given its recorded inputs and seeds, is bit-exact reproducible, reproducible within a stated tolerance, or not reproducible. The intent is that an interaction objective is optimized subject to its metered cost, so that the repertoire the system builds is one it can afford, and that every engagement it logs can be replayed and audited rather than merely reported.

These are engineering commitments, and honesty requires naming their status. The joules receipt is exact only for the parts of the pipeline that are instrumented, and is a bound or an estimate elsewhere; the replayability class is a property the substrate is built to guarantee, not a claim that every downstream component already honors it. Metering interaction in this way is, to our knowledge, not standard practice in the VLA literature, where energy and reproducibility are usually reported at the level of a whole training run if at all. Whether per-interaction metering changes what an interaction-maximizing agent learns is an open empirical question this report does not answer. The design is public. The Ambit repository is at github.com/dcharlot-physicalai-bmi/ambit^[10], and the MathGround substrate is documented with it^[11].

Table 1. Action versus interaction as an objective, and what metering adds. The right column states design intent, not measured outcomes.

ASPECT	VLA – ACTION / TASK-COMPLETION	VLI – INTERACTION, METERED (AMBIT)
What is maximized	success on named tasks	interaction with the world (curiosity / empowerment family)

What is shaped	trajectories to named goals	a repertoire over reachable situations
Readiness for the unnamed	incidental, if it appears at all	the direct object of the objective
Cost accounting	usually per training run, if reported	per interaction: a joules receipt
Reproducibility	rarely a per-decision property	per interaction: a replayability class

6. What a repertoire built by interaction buys

The reason to prefer an interaction objective is not that it completes named tasks better; it need not. It is that it produces the thing an order-following system lacks, namely competence in situations no order named. A system that has spent its training interacting has, before any command arrives, discovered what happens when it pushes a thing that resists, when a surface gives, when a grasp slips, when another agent moves into its path. Each such discovery is an entry in a repertoire, and the repertoire is what makes an unseen situation tractable, because an unseen situation is rarely wholly new; it is usually a recombination of interactions the system has already had. This is the sense in which maximizing interaction is, by construction, assembling the repertoire needed to act in any situation the system meets. It is the same argument the intrinsic-motivation literature makes for exploration^[5,6] and the empowerment literature makes for keeping options open^[7,8], transposed to the embodied, language-conditioned setting where the situations are physical contacts and the repertoire is motor and perceptual.

Two qualifications keep this from being a slogan. First, coverage is not free, which is why Section 5 prices it; an unbounded repertoire is neither affordable nor useful, and the metered objective is what keeps the accumulation disciplined. Second, a repertoire is not a guarantee. Readiness for the unseen is a matter of degree, and there will always be situations far enough outside the region of Figure 1 that no prior interaction helps. The claim is comparative, not absolute: a system whose objective was to interact will, on average, meet a wider band of the world prepared than a system whose objective was to complete the tasks it was shown. That comparative claim is the position; establishing it quantitatively is future work.

7. Position and discussion

It is worth being exact about what in this report is a position and what is demonstrated, because the two are easy to blur. Demonstrated, by others and cited here, are the following: VLA models transfer web knowledge to control and complete open-vocabulary tasks^[1,2]; concurrent perception and action within a single embodied agent is achievable^[3]; policies can improve from their own experience^[4]; curiosity and empowerment produce useful task-free behavior^[5,7]. A position, argued here and not proven, is that making interaction the training objective for embodied language-conditioned models is preferable to making task-completion the objective, and that doing so is what supplies readiness for the unnamed. A design commitment, stated but not benchmarked in this report, is that Ambit meters each interaction

with a joules receipt and a replayability class on MathGround. We do not claim that interaction-as-objective is novel; its ingredients are decades of prior art in intrinsic motivation, artificial curiosity, and empowerment, and Figure 1 is a restatement of those ideas, not a discovery. We do claim that stating the objective plainly for the VLA era, and pricing it per interaction, is a useful and under-explored framing. The most substantive open questions are empirical: whether an interaction objective in fact yields broader readiness than a task objective at equal compute, and whether per-interaction metering changes the repertoire that is learned. Neither is settled here.

8. Conclusion

A digital system with agency cannot only take orders, because the world it must act in is larger than any order it is given. This report advanced a research position, Vision-Language-Interaction, that makes interaction itself the objective, so that the repertoire needed to act in unseen situations is built by construction rather than hoped for as a byproduct of task training. The position is not new in its parts; it restates the intrinsic-motivation, curiosity, and empowerment traditions for the embodied, language-conditioned setting, and it is cited as such. What the lab adds is a commitment to ground and to meter: in *Ambit*, every interaction runs on MathGround and carries a joules receipt and a replayability class, so that a decision to engage the world is affordable and auditable. Whether this objective delivers the broader readiness it promises is an empirical question, and stating it precisely is the contribution this report intends to make.

References

1. A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, et al. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. arXiv:2307.15818, 2023.
2. M. J. Kim, K. Pertsch, S. Karamcheti, et al. OpenVLA: An Open-Source Vision-Language-Action Model. arXiv:2406.09246, 2024.
3. X. Liu, et al. VITA-E: Natural Embodied Interaction with Concurrent Seeing, Hearing, Speaking, and Acting. arXiv:2510.21817, 2025.
4. Physical Intelligence. $\pi^*0.6$: a VLA that Learns from Experience (RL with Experience and Corrections via Advantage-conditioned Policies, "Recap"). Technical disclosure, pi.website/blog/pistar06, 2025.
5. D. Pathak, P. Agrawal, A. A. Efros, T. Darrell. Curiosity-driven Exploration by Self-supervised Prediction. ICML 2017; arXiv:1705.05363.
6. J. Schmidhuber. Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990–2010). IEEE Trans. Autonomous Mental Development 2(3):230–247, 2010. doi:10.1109/TAMD.2010.2056368.
7. A. S. Klyubin, D. Polani, C. L. Nehaniv. Empowerment: A Universal Agent-Centric Measure of Control. Proc. IEEE Congress on Evolutionary Computation (CEC), pp. 128–135, 2005.
8. C. Salge, C. Glackin, D. Polani. Empowerment — An Introduction. In M. Prokopenko (ed.), Guided Self-Organization: Inception, Springer, pp. 67–114, 2014; arXiv:1310.1863.
9. P.-Y. Oudeyer, F. Kaplan. What is Intrinsic Motivation? A Typology of Computational Approaches. Frontiers in Neurobotics 1:6, 2007. doi:10.3389/neuro.12.006.2007.
10. D. J. Charlot. *Ambit*: Vision-Language-Interaction on a metered substrate. Source repository, github.com/dcharlot-physicalai-bmi/ambit, 2026.

11. The Charlot Lab. MathGround: a substrate attaching a joules receipt and a replayability class to each computation. Institute for Physical AI @ BMI, documented with the Ambit repository, 2026.

AI-use disclosure. Preparation of this report used a large language model (Claude, Anthropic) for drafting and editing text, organizing the reviewed literature, and preparing the figures and the interactive companion. Cited references were checked to resolve to their sources. The author reviewed the content and is solely responsible for it. Consistent with ICMJE, COPE, and IEEE guidance, the model is a tool and is not credited as an author.

Institute for Physical AI @ BMI · The
Charlot Lab
503 McKeever Rd, Arcola, TX 77583, USA
physicalai-bmi.org · contact@physicalai-
bmi.org

Technical Report TR-2026-16 · Preprint v1
© 2026 Institute for Physical AI @ BMI.
Released for open scholarly use. No proprietary or novel experimental results
are reported.