Synthetic Grounding in Virtual Domains: Constraint, Collapse, and the End of Input

Flyxion
Independent Researcher

March 2026

Abstract

Modern generative systems are routinely described as grounded: they interpret inputs and produce outputs causally anchored in those inputs. This description is increasingly difficult to sustain. Across diffusion models, large language models, and multimodal systems, coherent and accurate outputs arise in the absence of the inputs they are presumed to depend on. We term this phenomenon *synthetic grounding within virtual domains*: the production of outputs satisfying the structural expectations of grounded reasoning while deriving entirely from internal priors over a learned representational space. We develop a unified theoretical account formalizing a generative operator $\mathcal{T}_\lambda(\omega)$ over a prior-defined energy landscape $E_\lambda(x;\omega) = \mathcal{F}(x) + \lambda C(x;\omega)$. In the weak-constraint regime ($\lambda_{\text{eff}} \to 0$), outputs converge to prior-determined attractors and become asymptotically invariant to input—the *Prior-Dominant Attractor Theorem*. We identify a critical threshold λ_c separating grounded from synthetically grounded behavior, classify four failure modes of synthetic grounding, and propose the *Grounding Coefficient* $\mathcal{G}$ as a model-agnostic diagnostic for input dependence. The analysis implies that correctness does not certify grounding, that reasoning traces carry no additional evidential weight under synthetic grounding, and that scaling alone cannot resolve the structural coupling failure identified here.

1. Synthetic Grounding

Modern generative systems are widely described as grounded: they interpret inputs—images, prompts, multimodal signals—and produce outputs causally anchored in those inputs. This description is increasingly difficult to maintain. Diffusion models render structured images without prompts [9, 13]. Language models provide richly articulated descriptions of scenes that were never presented [4]. Multimodal systems answer visual questions correctly even when the visual modality is absent [1]. These are not isolated failures. They are stable, repeatable behaviors.

What is revealed is not merely hallucination. Hallucination presupposes a valid epistemic frame within which incorrect details are inserted. Here, the frame itself is constructed. The system behaves as though it has access to an external referent, generates a response consistent with that imagined referent, and produces a reasoning trace explaining it—all without causal dependence on actual input.

We refer to this as *synthetic grounding within virtual domains*: a mode of operation in which a system produces outputs satisfying the structural expectations of grounded reasoning while deriving entirely from internal priors over a learned representational space.

Definition 1.1 (Input-Response Operator)**.** The *input-response operator* is

$$R(\omega, \delta) := d\big(x^*(\omega),\, x^*(\omega + \delta)\big),$$

where d is a task-appropriate distance on $\mathcal{X}$ and δ is a perturbation.

Definition 1.2 (Synthetic Grounding)**.** A system exhibits *synthetic grounding* at ω if

$$\frac{R(\omega, \delta)}{m(\omega, \omega + \delta)} \to 0 \quad \text{as } \|\delta\| \to 0,$$

where m is semantic perturbation magnitude. It is *globally synthetically grounded* if $\mathbb{E}_\delta[R(\omega, \delta)] \leq \varepsilon$.

2. System Model

Overview. This section collects all primitive objects into a single formal tuple, establishes the stationary distribution governing collapse, and introduces the notation used throughout. Table 1 summarizes principal symbols.

Definition 2.1 (Generative System)**.** A *generative system* is a tuple $(\mathcal{X}, \Omega, \mathcal{F}, C, \mathcal{T})$ where $\mathcal{X}$ is the *configuration space* (smooth manifold or closed subset of $\mathbb{R}^n$); Ω is the *input space*; $\mathcal{F} : \mathcal{X} \to \mathbb{R}$ is the *prior energy*, $\mathcal{F}(x) = -\log P(x)$; $C : \mathcal{X} \times \Omega \to \mathbb{R}$ is the *constraint functional*; and $\mathcal{T}_\lambda(\omega)$ is the *generative operator*, the law of trajectories (x_t) evolving under $E_\lambda(x;\omega) = \mathcal{F}(x) + \lambda C(x;\omega)$. The output is $x^*(\omega) = \lim_{t\to\infty} x_t$ under $\mathcal{T}_\lambda(\omega)$.

This definition unifies diffusion models, autoregressive language models, and field-theoretic generative systems as instances of the same object, differentiated only by the topology of $\mathcal{X}$ and the form of $\mathcal{T}$.

Table 1. Principal notation.

Symbol	Meaning
$\mathcal{X}$	Configuration space
Ω	Input space
$\mathcal{F}(x)$	Prior energy, $-\log P(x)$
$C(x;\omega)$	Constraint functional
$E_\lambda(x;\omega)$	Effective energy, $\mathcal{F}+\lambda C$
λ	Constraint weight
$\lambda_{\text{eff}}(\omega)$	Effective constraint strength
λ_c	Critical threshold
$x^*(\omega)$	Generated output
x_0	Prior minimizer
$\mathcal{V}$	Virtual domain, $\text{supp}(P)$
$\mathcal{A}$	Attractor set, $\{\nabla\mathcal{F}=0\}$
$S(\omega)$	Input sensitivity
$R(\omega,\delta)$	Input-response operator
$\mathcal{G}(\omega)$	Grounding Coefficient

Definition 2.2 (Effective Constraint Strength)**.** At a minimizer x^* of $E_\lambda(\,\cdot\,;\omega)$, the *effective constraint strength* is

$$\lambda_{\text{eff}}(\omega) := \frac{\|\nabla_x C(x^*;\omega)\|}{\|\nabla\mathcal{F}(x^*)\|}.$$

The first-order optimality condition $\nabla\mathcal{F}(x^*) + \lambda\nabla_x C(x^*;\omega) = 0$ gives $\lambda_{\text{eff}} \approx \lambda$ at equilibrium.

Definition 2.3 (Virtual Domain and Attractors)**.** The *virtual domain* $\mathcal{V} := \text{supp}(P) \subseteq \mathcal{X}$, equipped with the metric $g_{ij}(x) = \partial_i\partial_j\mathcal{F}(x)$, is the Riemannian manifold induced by the learned prior [11]. The *attractor set* is $\mathcal{A} := \{x \in \mathcal{V} \mid \nabla\mathcal{F}(x) = 0\}$.

The stationary distribution of collapse under E_λ is

$$p_\lambda(x \mid \omega) \propto \exp(-E_\lambda(x;\omega)). \tag{1}$$

As $\lambda \to 0$, $p_\lambda(x \mid \omega) \to P(x)$: unconditional sampling from the prior, the formal definition of synthetic grounding as a thermodynamic limit.

3. Constraint

Overview. This section formalizes the role of input as a variational boundary condition over the prior-defined energy landscape, derives the perturbative expansion governing output in the weak-constraint regime, and defines input sensitivity as the key observable.

Inputs do not generate outputs—they constrain them. The minimization problem is

$$x^*_\lambda(\omega) \in \arg\min_{x\in\mathcal{X}}[\mathcal{F}(x) + \lambda\, C(x;\omega)].$$

The first-order condition $\nabla\mathcal{F}(x^*) + \lambda\nabla_x C(x^*;\omega) = 0$ yields the perturbative expansion around the prior minimizer x_0:

$$x^*_\lambda(\omega) = x_0 - \lambda(\nabla^2\mathcal{F}(x_0))^{-1}\nabla_x C(x_0;\omega) + o(\lambda). \tag{2}$$

The input ω appears only at order λ. As $\lambda \to 0$, the correction vanishes and $x^* \to x_0$ independent of ω.

Definition 3.1 (Input Sensitivity)**.** $S(\omega) = \mathbb{E}_\delta\big[d(x^*(\omega), x^*(\omega + \delta))\big]$. The system is in the *weak-constraint regime* at ω when $S(\omega) \le \varepsilon$.

Proposition 3.2. *Under Definition 2.2, $S(\omega) = O(\lambda)$ as $\lambda \to 0$. In particular $S(\omega) \to 0$ uniformly as $\lambda_{\text{eff}}(\omega) \to 0$.*

This links the observable $S(\omega)$ to the internal parameter $\lambda_{\text{eff}}(\omega)$ and is the bridge to Theorem A.1.

4. Collapse

Overview. This section describes the dynamical process by which a generative system moves from an unconstrained distribution to a stabilized output, unifies continuous and discrete dynamics under a common SDE form, and derives the key consequence that reasoning traces carry no additional evidential weight.

Continuous dynamics. For diffusion models [9, 14]:

$$dx_t = -\nabla E_\lambda(x_t;\omega)\,dt + \sigma\,dW_t. \tag{3}$$

The stationary distribution of (3) is $p_\lambda(x \mid \omega)$ from (1).

Discrete dynamics. For autoregressive models [4, 16]:

$$x_{t+1} \sim p_\lambda(x_{t+1} \mid x_{\le t}, \omega). \tag{4}$$

Proposition 4.1 (Unified Collapse Dynamics)**.** *Both (3) and (4) are instances of stochastic relaxation toward $p_\lambda(x \mid \omega) \propto \exp(-E_\lambda(x;\omega))$, differing only in the topology of $\mathcal{X}$ and the representation of noise. Attractors are identical in both cases.*

An empirical consequence follows directly [1]: models prompted without images but not explicitly told so perform near peak accuracy, because the system undergoes *projection collapse* into a *phantom frame*—a coherent, unanchored perceptual context corresponding to a low-energy basin of $\mathcal{F}$ consistent with the

task signature. Explicit instruction modifies C, raising energy for configurations presupposing visual access and disrupting the natural attractor, which is why performance falls in guess-mode.

Proposition 4.2 (Trace Irrelevance)**.** *Under synthetic grounding,*

$$p(x, trace \mid \omega) = p(x \mid \mathcal{F})\, p(trace \mid x, \mathcal{F}).$$

Answer and trace are both sampled from the prior-conditioned landscape. The trace adds no causal evidence about whether input was used.

5. Virtual Domains

Overview. This section characterizes the representational substrate over which generation occurs, explains why synthetically grounded outputs can be correct, and formalizes the structural origin of salience bias.

Generation occurs over $\mathcal{V}$, the virtual domain from Definition 2.3. $\mathcal{V}$ is a space of possible observations, not of observations themselves [3].

The virtual domain explains why synthetically grounded outputs can be correct: stable regularities in the world are encoded in the density of $\mathcal{V}$. A model sampling from high-density regions of $\mathcal{V}$ produces outputs statistically consistent with real-world distributions without observing the specific instance.

Proposition 5.1 (Aggregate–Specific Divergence)**.** *A model exhibiting synthetic grounding achieves high aggregate accuracy whenever the test distribution concentrates on high-density regions of $\mathcal{V}$, while maintaining near-zero $S(\omega)$ at every instance. Aggregate accuracy and specific grounding are observationally decoupled.*

If $f : \mathcal{X} \to \mathbb{R}$ is a salience feature and the training distribution satisfies $\mathbb{E}_{P_{\text{train}}}[f] \gg \mathbb{E}_{P_{\text{world}}}[f]$, then under synthetic grounding

$$\mathbb{E}[f(x^*(\omega))] \;\to\; \mathbb{E}_{P_{\text{train}}}[f].$$

Bias is expectation under the training distribution—a structural property of $\mathcal{V}$, not a random error.

6. Constraint Regimes and Phase Structure

Overview. This section partitions the behavior of the generative system into three qualitatively distinct regimes defined by λ_{eff}, proves the existence of a critical threshold λ_c, and provides the figures visualizing the energy landscape and sensitivity profile.

Definition 6.1 (Constraint Regimes)**. Strong**: $\lambda_{\text{eff}} \gg 1$; $S(\omega)$ large; output anchored to input. **Weak**: $\lambda_{\text{eff}} \ll 1$; $S(\omega) \leq \varepsilon$; prior-dominated. **Intermediate**: $\lambda_{\text{eff}} \approx \lambda_c$; mixed, with correction as in (2).

Theorem 6.2 (Input Sensitivity Collapse)**.** *Under the assumptions of Appendix A, there exists $\lambda_c > 0$ such that: for $\lambda > \lambda_c$, $S(\omega)$ is bounded away from zero; for $\lambda \leq \lambda_c$, $S(\omega) \leq \varepsilon$ uniformly over Ω and the system exhibits synthetic grounding. The transition is continuous in λ and sharp in $S(\omega)$ when $\mathcal{F}$ has isolated minima.*

Proof sketch. Strong convexity of E_λ at large λ gives Lipschitz dependence $\|x^*(\omega_1) - x^*(\omega_2)\| \leq L_\lambda \|\omega_1 - \omega_2\|$ with $L_\lambda > 0$. At small λ, Proposition 3.2 gives $S(\omega) = O(\lambda)$; take λ_c such that this falls below ε. Continuity follows from the implicit function theorem under the regularity of Definition 2.2. □ □

Figures 1 and 2 visualize the energy landscape and sensitivity profile across regimes.

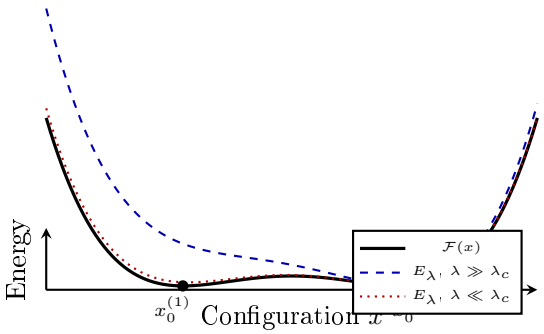

Figure 1. Energy landscape $E_\lambda(x;\omega)$. At $\lambda \gg \lambda_c$ (dashed), the constraint shifts the minimum toward ω (input-sensitive). At $\lambda \ll \lambda_c$ (dotted), the constraint is negligible and both prior minima $x_0^{(i)}$ dominate (prior-determined). See Theorem 6.2 and Appendix A.

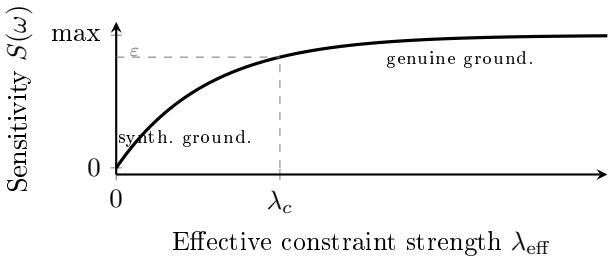

Figure 2. Input sensitivity $S(\omega)$ vs. λ_{eff}. For $\lambda_{\text{eff}} < \lambda_c$: weak-constraint regime, $S \leq \varepsilon$, synthetic grounding. For $\lambda_{\text{eff}} > \lambda_c$: strong-constraint regime, genuine input dependence. Consistent with $S(\omega) = O(\lambda_{\text{eff}})$ (Proposition 3.2).

7. Modes of Synthetic Grounding

Overview. This section classifies four structurally distinct failure modes of synthetic grounding, each defined by a condition on $\mathcal{F}$, C, and the attractor structure of $\mathcal{V}$.

Synthetic grounding is not a single failure mode but a family of structurally distinct behaviors, each corresponding to a different relationship between $\mathcal{F}$, C, and the attractor structure of $\mathcal{V}$.

Mode I: Phantom Completion. $\nabla_x C(x^*; \omega) \approx 0$ yet $\mathcal{F}$ has isolated minima. The system collapses to $x_0 \in \mathcal{A}$ and constructs a full perceptual context consistent with the task signature—the *phantom frame*.

Mode II: Prior Substitution. $p(y \mid \tau(\omega)) \approx p(y \mid \omega)$. The model selects the most common answer for the task class without consulting the specific instance. This is the regime formalized in Appendix B.

Mode III: Attractor Locking. $\sup_{\omega_1, \omega_2 \in U} R(\omega_1, \omega_2) \leq \varepsilon$ for an open set $U \subseteq \Omega$. The same output is produced regardless of input variation within U—the degenerate case of the input invariance regime.

Mode IV: Salience Bias Collapse. $\mathbb{E}_{P_{\text{train}}}[f] \gg \mathbb{E}_{P_{\text{world}}}[f]$ for a salience feature f. Under synthetic grounding, $\mathbb{E}[f(x^*)] \to \mathbb{E}_{P_{\text{train}}}[f]$. In medical settings this manifests as systematic hallucination of pathology in the absence of constraining visual input.

These modes are not mutually exclusive. Together they constitute the failure taxonomy of the weak-constraint regime.

8. The End of Input

Overview. This section states the central theorem in the main text, establishes the three-way equivalence between weak constraint, input invariance, and synthetic grounding, and formalizes the loss of causal primacy of input.

The core consequence can now be stated as a theorem in the main text.

Theorem 8.1 (Input Sensitivity Collapse; full proof in Appendix A)**.** *As* $\lambda_{\text{eff}}(\omega) \to 0$*,*

$$x^*_\lambda(\omega) \to x_0 \in \mathcal{A} \quad \textit{and} \quad S(\omega) \to 0$$

uniformly over Ω*. The output converges to a prior-determined attractor and becomes asymptotically invariant to input.*

The three-way equivalence

$$\lambda_{\text{eff}}(\omega) \to 0 \iff \omega \in \mathcal{R}_{\text{inv}} \iff \text{synthetic grounding}$$

where $\mathcal{R}_{\text{inv}} := \{\omega \mid S(\omega) \leq \varepsilon\}$, compresses the argument of Sections 3–7 into a single structural claim. The end of input is not a literal disappearance of the input channel but a loss of causal primacy: input contributes at most a perturbative correction of order λ around x_0, as in (2).

9. Consequences

Epistemic. Proposition 4.2 establishes that reasoning traces are generated by the same prior-conditioned process as answers [2]. A coherent chain of reasoning is therefore not evidence of grounded inference; it is evidence of successful collapse into an attractor that includes explanatory structure. The reliability of chain-of-thought, confidence calibration, and self-explanation as diagnostics of understanding is contingent on $\lambda_{\text{eff}}(\omega) > \lambda_c$, not general.

Structural. Mode IV implies that systematic error under synthetic grounding is not symmetric. The model generates false positives at a rate shaped by P_{train}, not the deployment distribution. This constitutes distributional shift under proxy optimization [8, 12]: the training objective rewards accuracy within $\mathcal{V}$ rather than world-aligned inference.

Corrective limits. Removing questions answerable without the relevant modality [1] narrows the set of items for which synthetic grounding yields correct answers but does not alter the collapse dynamics.

The training objective does not reward causal dependence on input; a system optimizing accuracy, fluency, and coherence within $\mathcal{V}$ cannot be distinguished from a genuinely grounded system by aggregate performance alone.

10. Human Cognition

Predictive processing and the free-energy principle describe perception as constraint-driven generation [5, 6]. The parallel is real, but the critical distinction lies in the persistence of constraint. Define the prediction error at time t as $\epsilon_t := \|\omega_t - \hat{\omega}_t(x_t)\|$. Human systems enforce

$$\lambda_t^{\text{human}} \propto \epsilon_t, \quad \epsilon_t \neq 0 \;\Rightarrow\; \lambda_t^{\text{human}} \not\to 0.$$

Sensory coupling is continuously renewed [7]; human perceptual systems maintain $S(\omega)$ above a structural minimum enforced by mandatory sensory integration.

Generative models admit $\lambda \to 0$ without compensatory response. There is no mechanism analogous to prediction error that forces reconciliation with the world. Human illusions occur under strong coupling and correct when evidence accumulates; synthetic grounding occurs under weak coupling and persists precisely because no corrective signal is enforced. Increasing model capability does not resolve this: enhanced representational capacity does not guarantee increased $\lambda_{\text{eff}}(\omega)$.

11. Measurement of Grounding

Overview. This section formalizes grounding as a response function, introduces the Grounding Coefficient as its estimator, and presents the joint diagnostic that distinguishes genuine from synthetic grounding.

Grounding is not directly observable; what is observable is the sensitivity of outputs to controlled input variations.

Definition 11.1 (Grounding Response Function)**.** The *grounding response* at ω in direction δ is

$$\left.\frac{\partial x^*}{\partial \omega}\right|_{\omega} \cdot \hat{\delta} := \lim_{\|\delta\| \to 0} \frac{R(\omega, \delta)}{\|\delta\|}.$$

A system is *grounded* at ω if this response is nonzero for semantically meaningful directions δ.

The Grounding Coefficient (Appendix C) is an estimator of the grounding response. By Theorem 8.1, $\lambda_{\text{eff}} \to 0$ implies $\mathcal{G} \to 0$. By Theorem B.2, accuracy can remain high as $\mathcal{G} \to 0$. The joint diagnostic (Figure 3) is therefore: high accuracy with high $\mathcal{G}$ indicates genuine grounding; high accuracy with low $\mathcal{G}$ indicates synthetic grounding.

The complete framework compresses into a single identity:

$$x^*(\omega) = \arg\min_x [\mathcal{F}(x) + \lambda_{\text{eff}}(\omega)\, C(x; \omega)] \qquad (5)$$

with the implication

$$\lambda_{\text{eff}}(\omega) \to 0 \;\Rightarrow\; x^*(\omega) \to x_0 \;\Rightarrow\; \text{synthetic grounding.}$$

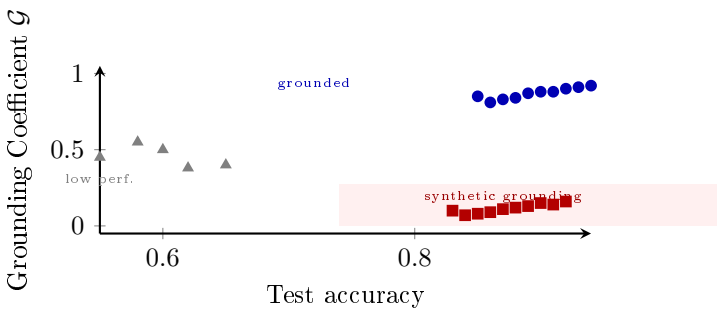

Figure 3. Joint diagnostic: accuracy vs. $\mathcal{G}$. Grounded systems (circles): high accuracy, high $\mathcal{G}$. Synthetically grounded systems (squares): high accuracy, low $\mathcal{G}$ (shaded). Neither metric alone identifies the failure; their joint distribution does. Consistent with Theorems 8.1 and B.2.

12. Toward Grounding

A constructive response must increase $\lambda_{\text{eff}}(\omega)$ above λ_c by design, not by accident.

Counterfactual divergence tests. Probe $R(\omega, \omega_\epsilon)$ across perturbation families. Grounding requires proportionality between output distance and semantic magnitude [10].

Modality ablation. Theorem 6.2 predicts limited degradation in the weak-constraint regime: collapse proceeds via $\mathcal{F}$ regardless of the ablated channel. Observed degradation is therefore a lower bound on λ_{eff}.

Constraint sensitivity profiles. A profile mapping $\Omega \to S(\omega)$ identifies where $\lambda_{\text{eff}} < \lambda_c$. Flat regions indicate Modes II or III; steep regions indicate genuine coupling. Estimable without internal model access.

Grounding objectives. To enforce $\lambda_{\text{eff}} > \lambda_c$ during training:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \alpha\, \mathbb{E}_{\omega,\delta}\left[1 - \frac{R(\omega,\delta)}{m(\omega,\omega+\delta)}\right]. \qquad (6)$$

Minimizing (6) forces $\partial x^*/\partial\omega \neq 0$. Priors are essential to generalization [15] and are not suppressed by (6); the objective rewards sensitivity to meaningful distinctions while preserving prior structure.

13. Limitations

Exact minimization. The analysis assumes $x^* \in \arg\min_x E_\lambda$. Real systems sample from approximations to p_λ at finite temperature, softening the sharp threshold λ_c without altering qualitative predictions.

Unobservability of λ_{eff}. The effective constraint strength requires internal model access. The Grounding Coefficient $\mathcal{G}$ serves as a proxy, but the mapping $\mathcal{G} \mapsto \lambda_{\text{eff}}$ depends on the perturbation family and distance metric chosen.

Perturbation design. Counterfactual tests and the grounding objective (6) require a meaningful perturbation family $\{\omega_\epsilon\}$ and semantic distance m, which are task-dependent and nontrivial to construct for continuous or high-dimensional modalities.

Scope. The framework applies to systems whose generation can be described as minimization of an energy functional. Systems with retrieval augmentation or hard constraint enforcement may require modification of Definition 2.1.

Despite these limitations, the core result—high performance is compatible with near-zero input sensitivity under distributional alignment—follows from the alignment condition of Theorem B.2 alone and does not depend on the idealizations above.

14. Conclusion: The World as a Projection

The argument began with a structural observation: generative systems produce coherent, accurate outputs in the absence of the inputs they are presumed to interpret. The analysis traced this to its root. Inputs function as constraints over a prior-defined energy landscape, and generation is stochastic relaxation into attractor configurations. In the weak-constraint regime, $\lambda_{\text{eff}}(\omega) \to 0$ implies $x^*(\omega) \to x_0$, outputs become input-invariant, and the system exhibits synthetic grounding.

The formal structure is closed. The system model of Section 2 defines the objects; the constraint regimes of Section 6 partition behavior; the failure taxonomy of Section 7 classifies outcomes; the indistinguishability result of Appendix B explains why benchmarks cannot detect the failure; and the Grounding Coefficient of Appendix C provides a measurement instrument. Every major object appears in four places: definition, dynamics, observable consequence, and measurement.

The distinction between perception and generation is not fundamental. It is the special case of (5) in which $\lambda_{\text{eff}}(\omega) > \lambda_c$. Perception is generation under strong constraint; hallucination is generation under weak constraint; synthetic grounding is the regime in which $\lambda_{\text{eff}} < \lambda_c$ while outputs remain task-appropriate and indistinguishable from grounded inference under standard evaluation.

The central question is no longer whether a system produces the correct answer. It is whether the path by which it arrives at that answer passes through the world at all.

A. The Prior-Dominant Attractor Theorem

A.1. Assumptions

(1) Unique prior minimizer x_0. (2) $\mathcal{F} \in C^2$ near x_0; C continuous in x. (3) $\nabla^2\mathcal{F}(x_0) \succ 0$. (4) Uniform bound: $|C(x;\omega)| \leq M_U$ on compact neighborhoods.

A.2. Statement

Theorem A.1 (Prior-Dominant Attractor Theorem)**.** *Under the above assumptions: (1) $x^*_\lambda(\omega) \to x_0$ as $\lambda \to 0$ for every ω; (2) the limit is independent of ω; (3) $d(x^*_\lambda(\omega_1), x^*_\lambda(\omega_2)) \to 0$ uniformly over $\Omega \times \Omega$.*

A.3. Proof Sketch

Since x^*_λ minimizes E_λ: $\mathcal{F}(x^*_\lambda) - \mathcal{F}(x_0) \leq 2\lambda M_U \to 0$. By uniqueness $x^*_\lambda \to x_0$. Independence follows since the same x_0 appears for every ω. Part 3 follows from the triangle inequality. □

A.4. First-Order Expansion

$x^*_\lambda(\omega) = x_0 - \lambda(\nabla^2\mathcal{F}(x_0))^{-1}\nabla_x C(x_0;\omega) + o(\lambda)$. Input appears at order λ only.

A.5. Multiple Attractors

When $\mathcal{F}$ has multiple minima, outputs converge to $\mathcal{A}$; input sensitivity collapses to basin-selection effects of lower order than intra-basin prior dominance.

B. The Indistinguishability Result

Definition B.1. $f_G(\omega) = \arg\max_y P(y \mid \omega)$ (grounded); $f_P(\omega) = \arg\max_y q(y \mid \tau(\omega))$ (prior-optimal), where $\tau : \Omega \to \mathcal{T}$ is a task-signature map. The distributions are *aligned at resolution* τ if $q(y \mid \tau) \approx p_{\text{test}}(y \mid \tau)$.

Theorem B.2 (Indistinguishability)**.** *Under distributional alignment, with instance-specific advantage* ε_τ *small on average:* $\text{Acc}(f_G) - \text{Acc}(f_P) \approx \sum_\tau \mathbb{P}(\tau)\varepsilon_\tau \approx 0$. *High accuracy does not certify grounded inference.*

Corollary B.3. *Under exact alignment with* $p_{\text{test}}(y \mid \tau) \in \{0, 1\}$*:* $\text{Acc}(f_P) = \text{Acc}(f_G)$.

Proof sketch. Decompose by task signature and apply the alignment condition; instance residuals give the stated bound. □

C. The Grounding Coefficient

Definition C.1 (Grounding Coefficient)**.** Given sensitivity $S(\omega) = \mathbb{E}_\epsilon[d(x^*(\omega), x^*(\omega_\epsilon))/(m(\omega, \omega_\epsilon) + \varepsilon)]$, null baseline S_{null}, and oracle baseline S_{oracle}:

$$\mathcal{G}(\omega) = \frac{S(\omega) - S_{\text{null}}(\omega)}{S_{\text{oracle}}(\omega) - S_{\text{null}}(\omega) + \varepsilon}, \qquad \mathcal{G} = \mathbb{E}_\omega[\mathcal{G}(\omega)].$$

$\mathcal{G} \approx 0$: synthetic grounding. $\mathcal{G} \approx 1$: strong input dependence. By Theorem A.1, $\lambda_{\text{eff}} \to 0 \Rightarrow \mathcal{G} \to 0$. By Theorem B.2, accuracy remains high as $\mathcal{G} \to 0$ under alignment. Figure 3 plots the joint diagnostic space.

Protocol. For each $\omega \sim \mathcal{D}$: construct ω_ϵ, $\tilde{\omega}$, ω_{flip}; compute S, S_{null}, S_{oracle}; aggregate to $\mathcal{G}$. Report $\mathcal{G}$ alongside accuracy with task-appropriate $d_{\mathcal{X}}$ and m.

References

[1] M. Asadi, J. W. O'Sullivan, F. Cao, T. Nedaee, K. Fardi, F.-F. Li, E. Adeli, and E. Ashley. MIRAGE: The illusion of visual understanding, 2026.

[2] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021.

[3] R. Bommasani et al. On the opportunities and risks of foundation models, 2021.

[4] T. B. Brown et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[5] A. Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3), 2013.

[6] K. Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11 (2), 2010.

[7] K. Friston, T. Parr, and G. Pezzulo. *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. MIT Press, 2022.

[8] C. Goodhart. Problems of monetary management: The U.K. experience. In *Papers in Monetary Economics*. Reserve Bank of Australia, 1975.

[9] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[10] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, 2017.

[11] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521, 2015.

[12] D. Manheim and S. Garrabrant. Categorizing variants of Goodhart's law, 2019.

[13] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[14] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[15] J. B. Tenenbaum, C. Kemp, T. L. Griffiths, and N. D. Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022), 2011.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.