Activation Manifolds as Admissibility Fields: Reachability Geometry in Large Language Models

Flyxion

Independent Research

June 2026

Abstract

Three independent empirical results from June 2026 — normalization-revealed activation geometry [22], ICA-recovered non-Gaussian directional structure [23], and trajectory-field self-distillation in world models [30] — converge on a single geometric picture that has not previously been articulated in a unified form. We argue that what these three programs are collectively discovering in activation space is not feature geography but an emergent shadow of a deeper constraint topology, which we formalize as an *admissibility field* $A : \mathcal{M} \to \mathbb{R}_{\geq 0}$ defined over the activation manifold. This object is not imported as a prior theoretical framework; it emerges from the empirical observations as the minimal geometric entity capable of simultaneously explaining all three findings. The central inversion the evidence supports is: *future structure determines representation*, reversing the standard assumption that representation determines behavior. We formalize this as the *Reachability Determines Representation* principle, prove a formal equivalence relation over activation states, and derive from it a unified field hypothesis in which activation clusters, ICA directions, effective receptive

fields, and velocity fields in trajectory space all appear as different measurements of the same underlying object $A(x)$. The paper also establishes a structural correspondence between the three empirical programs and the three foundational layers of modern differential geometry: topology (which asks what is connected), differential structure (which asks what directions exist locally), and dynamics (which asks how trajectories evolve). This triad maps onto Liu's activation graph, ICALens's independent components, and WMSD's learned velocity fields almost perfectly, suggesting that the architecture of modern geometry was, in a precise sense, the right language for activation space all along.

Keywords: activation geometry, admissibility fields, reachability, mechanistic interpretability, independent component analysis, world models, representation learning, manifold theory, constraint topology

Contents

1. Introduction: Three Convergences

Theoretical frameworks in science earn their keep not by fitting a single observation but by revealing that several apparently unrelated observations are, in fact, measurements of the same underlying object seen from different angles. When this happens — when a minimal formal entity renders transparent a cluster of empirical findings that had seemed to require separate explanations — the result is not merely a new theory but a kind of perceptual reorganization. What had looked like a landscape of disconnected facts becomes a topographic map of a single terrain.

This paper argues that something of this kind is occurring right now in the mechanistic interpretability of large language models. Three independent empirical programs, developed in near-simultaneity during the first half of 2026, have arrived at structurally identical conclusions by entirely different routes. The first, Liu's normalization study [22], demonstrates that once the magnitude of residual-stream activations is discarded via L2 normalization, the resulting points on the unit hypersphere organize into coherent connected components — and that those components correspond not to semantic topics such as "mammals" or "countries" but to structural and procedural constraints such as question-and-answer formatting, copyright boilerplate, patent section headings, LaTeX syntax, and programming language import blocks. The second, the ICALens study by Liu and Han [23], demonstrates that the same activation space, when subjected to row-normalization followed by whitening and independent component analysis, reveals directional structure characterized by non-Gaussian, heavy-tailed projection distributions — and that these statistically exceptional directions correspond to modes of textual constraint rather than to object categories or named concepts. The third, the World Model Self-Distillation paper by Stapf et al. [30], demonstrates that an instruction-conditioned video world model trained on a rich teacher signal containing step-by-step solution trajectories can, through a process of self-distillation combined with reinforcement learning and vision-language model feedback, learn to match or exceed the teacher's performance using only high-level task specifications, with additional transfer to robotic tasks — and that the right mathematical formalization of what the student is learning is not a set of states or a policy function but a *velocity field over possibility space*, whose fundamental stability properties are governed by local matching of continuation dynamics rather than by state-level imitation.

Each of these results is striking in isolation. Together they are extraordinary. What they share, as we will argue in careful detail, is a single structural claim:

that the computational geometry of learned representations is organized primarily around *what can happen next* rather than around *what is present now*. The activation manifold appears to be partitioned not into categories of objects but into equivalence classes of futures — regions where different tokens, contexts, and computational states become interchangeable because they license nearly the same continuation possibilities. The relevant notion of similarity, in other words, is not lexical or even semantic in the traditional sense; it is *reachability-theoretic*.

The formal object we introduce to unify these observations is the *admissibility field* $A : \mathcal{M} \to \mathbb{R}_{\geq 0}$, where $\mathcal{M}$ is the activation manifold. Informally, $A(x)$ measures, at activation state x, how much admissible future trajectory lies ahead — the measure of the reachable future set weighted by a characteristic function of valid continuations. This object is deliberately not introduced as a presupposed theoretical framework. We introduce it only in Section 4, after establishing the empirical anomalies in Sections 2 and 3, and we give its full formal synthesis only in Section 11, after deriving its role in each of the intermediate sections independently. The strategy throughout is: find the observations, identify the constraints each places on any explanatory structure, and determine the minimal formal entity that satisfies all of them. The admissibility field is our candidate for that entity.

A note on intellectual genealogy. The synthesis we are proposing does not emerge from nowhere. It draws on a long tradition in manifold learning [32, 27, 5], information geometry [1], percolation theory [31, 24], and the emerging field of mechanistic interpretability [25, 8, 4, 7]. It connects to the theoretical program in world models [14, 20, 30] and to the reachability-theoretic ontology developed across several prior monographs in this series [12, 11, 9, 10]. The bibliography at the end of this paper is organized deliberately to reflect the logical genealogy of the argument: geometry and topology first, interpretability second, dynamics third, reachability fourth. This ordering is not incidental; it mirrors the structure of the paper itself and is intended to make clear that the admissibility field emerges from the empirical literature rather than being imposed upon it.

The structure of the paper is as follows. Section 2 formalizes Liu's normalization procedure as a projection operation in the sense of differential geometry and derives its implications for the structure of angular similarity. Section 3 examines the empirical clustering results and argues that the content of the clusters — procedural and structural constraints rather than semantic topics — constitutes a fundamental anomaly that forces a reexamination of what activation representations encode. Section 4 introduces the Future Equivalence Principle and the Reachability Determines Representation inversion, which together provide the theoretical bridge between the clustering observations and the admissibility field formalism. Section 5

analyzes the topology of the activation graph, reinterpreting the giant component result and the power-law component spectrum in terms of the level-set topology of $A(x)$. Section 6 provides the bridge between graph-theoretic and differential descriptions, establishing the correspondence between connected components, ICA directions, and velocity fields as instantiations of the topology-differential structure-dynamics triad. Section 7 formalizes ICALens's key technical innovations — non-Gaussianity as constraint selectivity, effective receptive field as correlation length, ICA versus SAE — within the admissibility geometry framework. Section 8 derives the WMSD velocity-field formalism and re-reads its key stability theorem as a statement about local admissibility agreement controlling global reachability. Section 9 formalizes distillation as an admissibility-preserving projection and establishes the Task Sufficiency Projection Principle. Section 10 consolidates the ontological inversion across all three papers, arguing that it is an empirical consequence rather than a philosophical assertion. Section 11 assembles the Admissibility Field Hypothesis in its full formal version and identifies the open mathematical questions the framework raises. Section 12 concludes.

2. Normalization as Projection: The Hyperspherical Turn

2.1. The Structure of Residual Activations

Let $h \in \mathbb{R}^d$ be a residual-stream activation of a transformer at some layer. For the Gemma model analyzed in [22], $d = 2304$. Each token processed by the model at each layer generates one such vector. If we collect n tokens, we obtain a point cloud $\{h_1, h_2, \ldots, h_n\} \subset \mathbb{R}^d$. In Liu's experiment $n \approx 10^6$.

In its raw form, each activation h_i has both a direction and a magnitude:

$$h_i = \|h_i\| \cdot \hat{h}_i, \qquad \hat{h}_i = \frac{h_i}{\|h_i\|}. \tag{1}$$

The full information content of h_i is therefore encoded in the pair $(\|h_i\|, \hat{h}_i) \in \mathbb{R}_{>0} \times \mathbb{S}^{d-1}$. These two components are in principle independent, and they carry different kinds of information about the model's computational state. Magnitude can reflect confidence, salience, or energy-like properties of the current processing context; direction encodes the relational and positional structure within the high-dimensional feature space.

2.2. Normalization as a Canonical Quotient Map

L2 normalization discards magnitude and retains only direction. Formally, it is the map

$$\pi_{\mathbb{S}} : \mathbb{R}^d \setminus \{0\} \longrightarrow \mathbb{S}^{d-1}, \qquad \pi_{\mathbb{S}}(h) = \frac{h}{\|h\|}. \tag{2}$$

This map has a clean algebraic description. The group $\mathbb{R}_{>0}$ of positive real numbers acts on $\mathbb{R}^d \setminus \{0\}$ by scalar multiplication: $\lambda \cdot h = \lambda h$ for $\lambda > 0$. Two vectors are in the same orbit under this action if and only if they have the same direction, i.e., $h \sim h'$ iff $h' = \lambda h$ for some $\lambda > 0$. The map $\pi_{\mathbb{S}}$ is precisely the quotient map for this equivalence relation, and the quotient space is

$$\big(\mathbb{R}^d \setminus \{0\}\big)/\mathbb{R}_{>0} \;\cong\; \mathbb{S}^{d-1}. \tag{3}$$

This algebraic fact has a significant geometric implication: normalization is not an arbitrary preprocessing step but a geometrically natural operation corresponding to the identification of scale-equivalent vectors. It reduces the dimensionality of the relevant structure by one (the scale degree of freedom) while leaving the directional degrees of freedom intact.

2.3. Angular Similarity as the Natural Metric

Once all activations lie on $\mathbb{S}^{d-1}$, the natural metric is angular. For two normalized activations $\hat{h}_i, \hat{h}_j \in \mathbb{S}^{d-1}$, the squared Euclidean distance is

$$\left\|\hat{h}_i - \hat{h}_j\right\|^2 = \left\|\hat{h}_i\right\|^2 + \left\|\hat{h}_j\right\|^2 - 2\left\langle \hat{h}_i,\, \hat{h}_j \right\rangle = 2 - 2\left\langle \hat{h}_i,\, \hat{h}_j \right\rangle = 2\big(1 - \cos(h_i, h_j)\big), \tag{4}$$

where $\cos(h_i, h_j) = \left\langle \hat{h}_i,\, \hat{h}_j \right\rangle$ is the cosine similarity. This means that Euclidean proximity on the hypersphere is equivalent to cosine similarity, and that the graph connectivity threshold θ in Liu's construction (two activations are connected when their cosine similarity exceeds θ) corresponds exactly to a geodesic ball condition on the sphere.

The choice to use angular rather than Euclidean distance is not merely aesthetic. It reflects a deep assumption about what matters computationally. The RMSNorm operation used in modern transformers [34] normalizes activations before applying learned weight matrices, which means that the downstream computation at each layer effectively operates on the normalized direction $\hat{h}_i$ rather than the raw vector h_i. The downstream layers are, in a precise functional sense, insensitive to magnitude and responsive only to direction. This is why [22] argues that normalization makes activation geometry "imaginable": it aligns the representation

space with the space that the computation is actually using.

2.4. The Astronomical Analogy and Degree-of-Freedom Reduction

There is a useful analogy from astronomy. When we observe the night sky, stars appear as points of light that vary enormously in apparent brightness. Attempting to detect constellations — geometric patterns of angular positions — in the presence of large brightness variation is difficult: a bright nearby star and a dim distant star may be spatially close on the celestial sphere but enormously different in apparent brightness. If one were to use raw Euclidean coordinates in 3-space rather than angular coordinates on the sphere, the positional structure would be obscured by scale variation.

Normalization performs precisely the same operation as projecting stars onto the celestial sphere: it discards the radial degree of freedom and reveals the angular geometry. The constellations become visible not because new information has been added but because an irrelevant degree of freedom has been removed. This is the general principle at work. Many representation-learning techniques [5, 32] can be understood as collapsing irrelevant fibers of a projection in order to expose the geometry that was present all along.

2.5. First Structural Observation

We have established the first structural observation of the paper: normalization is a geometrically natural projection that collapses the scale fiber $\mathbb{R}_{>0}$ over each direction, moving activations from $\mathbb{R}^d \setminus \{0\}$ to $\mathbb{S}^{d-1}$ and replacing Euclidean distance with angular similarity. Crucially, this operation aligns the representation with the downstream computation, since RMSNorm-based architectures are functionally insensitive to scale. The surprising empirical result that this projection *reveals* organized geometric structure rather than destroying it is the fact that demands explanation. We turn to that structure next.

3. Constraint Clusters, Not Concept Clusters

3.1. The Graph Construction

With normalized activations $\{\hat{h}_1, \ldots, \hat{h}_n\} \subset \mathbb{S}^{d-1}$ in hand, [22] constructs the graph

$$G_\theta = (V, E_\theta), \qquad (i,j) \in E_\theta \iff \left\langle \hat{h}_i, \hat{h}_j \right\rangle \geq \theta \tag{5}$$

for a fixed cosine similarity threshold θ. The vertices V are the million normalized activations; two vertices are connected by an edge whenever the angle between the corresponding directions falls below the threshold. Connected components of G_θ are then the maximal sets of activations that are mutually reachable through chains of high-similarity steps.

The central empirical result is the composition of the medium-sized and high-ranked components. We pause to state this carefully, because the precise content of the clusters is the observation that drives the entire theoretical argument.

3.2. The Central Anomaly

The naïve expectation, grounded in decades of word embedding and language model analysis, is that activation clusters would correspond to semantic topics. One expects to find a "mammals" cluster, a "countries" cluster, an "emotions" cluster, a "mathematical concepts" cluster. This expectation is deeply entrenched because it aligns with the distributional hypothesis in linguistics [28] and with the abundant evidence that language models encode semantic relationships in their representations.

What [22] finds is entirely different. The medium-sized connected components, which are the most informative (the giant component, as we will analyze in Section 5, is too large to be interpretable as a single cluster; the isolated singletons are too small), correspond to organizational and procedural structures of text:

The second-ranked cluster consists entirely of activations corresponding to the colon character in the pattern "Q:", i.e., the boundary between a question header and its content in question-and-answer formatted text. The other significant clusters correspond to copyright notice headers, patent section headings of the form "Field of the Invention" and similar, legal citation structures, software license boilerplate of the form "GNU General Public License", LaTeX mathematical notation delimiters and structural commands, JSON escape sequences and structural punctuation, Go language indentation patterns, and Java import block headers.

Let us be precise about why this is anomalous. These are not semantic topics. They share no common subject matter. A patent heading has nothing in common with a JSON escape sequence in terms of what it *refers to*. A copyright notice and a LaTeX fragment do not co-occur in documents, do not describe related concepts, and do not involve similar vocabulary. If semantic similarity were driving the clustering, we would expect such items to appear in entirely different parts of the activation space.

What they share is something else entirely: they are all located at boundaries into *constrained continuation regimes*. Each of these textual structures is a gateway

into a region of text space where the next tokens are highly predictable and the range of admissible continuations is severely restricted.

3.3. The Rank-2 Result in Detail

The second-ranked cluster — 276 activations all corresponding to the colon after "Q" — deserves especially careful analysis. The colon ":" is an extremely common character that occurs in thousands of different textual contexts: after labels in YAML files, in URL structures, in time expressions, after Greek citations, after ratios, in emoji, in mathematical notation, and in ordinary English prose. The fact that cosine similarity in normalized activation space selects out precisely and only the subset of colons that appear in question-and-answer context is not trivially explained by either the token's lexical identity or its syntactic role.

The 276 instances of "Q:" in the dataset come from many different documents, many different topics, many different authors, many different styles. They share precisely this: after "Q:", the model is in a highly specific computational state. The continuation is constrained to be a question or question-like text. The format that follows is expected to have specific structural properties. The future trajectory of the text, in a precise mathematical sense, is more constrained than it is for most other contexts. The activation for the colon after "Q" is not encoding the identity of the colon. It is encoding a location in *future-space*: the state of being at the entrance to a constrained continuation manifold.

Observation 3.1. The Rank-2 cluster in Liu's activation graph demonstrates that the network has learned activation states that are not attached to token identity but to the structure of the future continuation they initiate. The token ":" merely serves as an entrance into a region of state space; the activation encodes the properties of that region, not the properties of the token.

This observation is not limited to the Rank-2 cluster. It applies uniformly to the identified components. The activation corresponding to "Field of the Invention" is useful to the model not because it must identify patent headings as a lexical category but because, from that activation state, the admissible continuations are a very particular set: technical descriptions of inventions, problem statements, prior art summaries, legal claims, and similar patent-register prose. The activation corresponding to "import (" in Go source code is at a boundary into a structured block where the admissible tokens are package identifiers and structural delimiters. The activation corresponding to copyright boilerplate initiates a structured sequence whose continuation is almost entirely predictable.

3.4. Process Attractors versus Noun Clusters

The distinction being drawn here has a natural formulation in dynamical systems language. Classical attractor analysis in dynamical systems identifies regions of state space toward which trajectories converge under the governing dynamics. We can analogize: the activation clusters discovered by [22] are regions of activation space that are not distinguished by the object they represent but by the dynamic they initiate. They are process attractors rather than noun clusters.

A noun cluster would be a region of activation space that all activations for "dog"-related tokens converge to because dogs share common features. A process attractor is a region that activations converge to whenever the text enters a particular mode of structured continuation, regardless of the subject matter being discussed. The Q/A cluster is a process attractor for the beginning-of-question state; the patent-heading cluster is a process attractor for the beginning-of-patent-section state; the copyright cluster is a process attractor for the beginning-of-boilerplate state.

3.5. Second Structural Observation

This gives us the second structural observation: the geometric structure revealed by normalization is organized around continuation constraints rather than denotational categories. Activations cluster together not because they refer to similar things but because they initiate similar kinds of future trajectories. This is our central empirical anomaly, and it demands a theoretical account. In the next section we provide one.

4. The Future Equivalence Principle and the Admissibility Field

4.1. Why Cosine Similarity Is the Wrong Metric

The graph G_θ uses cosine similarity as its proximity criterion. This is a natural choice given that the activations have been projected onto the hypersphere, and it is the criterion that reveals the clustering structure described in Section 3. But the empirical content of the clusters suggests that cosine similarity is not the fundamental similarity relation at work; it is, rather, a proxy for a deeper relation.

To see this, consider the following thought experiment. Suppose we have two activations a_1 and a_2 that are far apart in angular distance on $\mathbb{S}^{d-1}$ but that both occur at the entrance to patent-heading continuation regimes. Standard cosine similarity would classify them as dissimilar. But from the perspective of what the

model will do next — what futures are available — they are functionally identical. Conversely, suppose a_1 and a_2 are cosine-similar in activation space but one occurs at a patent heading and one occurs in a general prose context. Cosine similarity would classify them as similar, but they license entirely different continuation manifolds.

The correct similarity relation is not determined by the current geometry of the activation vectors. It is determined by the geometry of the futures they generate. This motivates the central formal construction of the paper.

4.2. The Reachability Operator

Definition 4.1 (Reachability Operator)**.** Let $a \in \mathcal{M}$ be an activation state, where $\mathcal{M} \subset \mathbb{S}^{d-1}$ is the normalized activation manifold. Let $\tau = (a_{t_0}, a_{t_1}, \ldots, a_{t_k})$ denote a future activation trajectory beginning at $a_{t_0} = a$. The *reachability operator* at a is defined as

$$R(a) = \mathcal{L}(\tau \mid a), \tag{6}$$

the probability distribution over future activation trajectories τ conditioned on the current state a. Here $\mathcal{L}(\cdot \mid a)$ denotes the law of the trajectory process induced by the model's continuation dynamics starting from a.

Informally, $R(a)$ is the distribution over futures accessible from a. It encodes not merely the single most likely continuation but the entire manifold of continuation possibilities, weighted by their probability under the model's learned dynamics.

4.3. Future Proximity and Exact Equivalence

Definition 4.2 (ε-Future Proximity)**.** Let $D(\cdot, \cdot)$ be a distance on the space of trajectory distributions (for concreteness, one may take D to be the Wasserstein-2 distance or the KL divergence). For $\varepsilon \geq 0$, define the *ε-future proximity relation* on $\mathcal{M}$ by

$$a_1 \approx_R^{\varepsilon} a_2 \quad \Longleftrightarrow \quad D\big(R(a_1),\, R(a_2)\big) \leq \varepsilon. \tag{7}$$

In the limit $\varepsilon \to 0$ this recovers exact future equivalence: $a_1 \sim_R a_2 \iff R(a_1) = R(a_2)$.

Remark 4.3 (Transitivity). For $\varepsilon > 0$, the relation $\approx_R^{\varepsilon}$ is reflexive and symmetric but generally *not transitive*: one can have $a_1 \approx_R^{\varepsilon} a_2$ and $a_2 \approx_R^{\varepsilon} a_3$ without $a_1 \approx_R^{\varepsilon} a_3$, since the two balls of radius ε may overlap at a_2 while being separated at a_1 and a_3. The relation is therefore a *uniform tolerance structure* (a proximity relation in the sense of Zeeman) rather than a classical equivalence relation. Genuine equivalence classes partition $\mathcal{M}$ only in the exact limit $\varepsilon = 0$, yielding the quotient

$\mathcal{M}/\sim_R$ whose points are maximal future equivalents. Throughout the paper, "equivalence class" refers to this exact-limit partition; for finite ε we speak of *proximity neighborhoods*.

Future Equivalence Principle. Two activation states are functionally equivalent to the extent that they preserve the same set of admissible continuations. The natural similarity structure on activation states is future proximity $\approx_R^\varepsilon$, which recovers a genuine equivalence relation $\sim_R$ only in the exact limit $\varepsilon \to 0$. Cosine similarity is a proxy for this deeper structure, not the fundamental metric.

The motivation for this principle is the following informal but compelling argument. A transformer activation a is useful to the model exclusively insofar as it contributes to the computation of subsequent activations and ultimately to the probability distribution over next tokens. If two activations a_1 and a_2 induce nearly identical distributions over future activation trajectories — if $D(R(a_1), R(a_2)) \leq \varepsilon$ for small ε — then no downstream observer restricted to observing the model's behavior can distinguish between them: they produce the same probability distributions over any observable output. They are, in the most operationally meaningful sense, the same state.

Conversely, if $D(R(a_1), R(a_2))$ is large, then a_1 and a_2 generate observably different futures, even if they happen to be close in activation space by cosine similarity. Two states that are geometrically proximate but functionally distant — like a colon in prose versus a colon in "Q:" — represent different computational situations despite their superficial angular similarity.

4.4. The Reachability Determines Representation Conjecture

The Future Equivalence Principle gives us a refined notion of similarity. But it also opens the door to a much stronger claim. The empirical evidence from Section 3 suggests not merely that reachability *should* be used as a similarity criterion, but that the network has *learned* to organize its representation space according to future proximity classes. That is, the network's geometry has been shaped by its own future-prediction task in such a way that future-proximate states end up geometrically proximate.

An important clarification is required before stating this claim. The evidence in Section 3 establishes primarily that geometrically clustered states share similar continuation structures: clustering $\Rightarrow$ similar futures. The claim in the other direction — similar futures $\Rightarrow$ geometric clustering — is the stronger, causal claim

that the following conjecture asserts. The evidence motivates the conjecture but does not yet establish it.

Reachability Determines Representation (RDR Conjecture). If $R(a_1) \approx R(a_2)$, then a_1 and a_2 tend to become geometrically close in the learned activation space. Formally:

$$D\big(R(a_1), R(a_2)\big) \leq \varepsilon \quad \Rightarrow \quad \left\| \hat{h}_{a_1} - \hat{h}_{a_2} \right\| \leq f(\varepsilon), \tag{8}$$

where f is a monotone increasing function with $f(0) = 0$ that reflects the degree to which training geometry tracks reachability geometry.

Falsification criterion. The RDR conjecture is falsified if two activation states with nearly identical continuation distributions ($D(R(a_1), R(a_2)) \leq \varepsilon$ for small ε) are found to be systematically distant in the activation geometry. The existence of such pairs, if common, would mean that reachability similarity is necessary but not sufficient for geometric proximity, and some other factor determines the representational distance.

This conjecture inverts the standard assumed causal direction in representation learning. The traditional view is:

$$\text{Representation} \longrightarrow \text{Behavior}. \tag{9}$$

The token has a feature vector; that feature vector, through the transformer's computations, determines the probability distribution over next tokens. Representation is prior; behavior is derived.

The RDR conjecture reverses this:

$$\text{Future Structure} \longrightarrow \text{Representation}. \tag{10}$$

The relevant structure of the future continuation manifold — what is reachable, what is constrained, what is opened or closed by the current state — determines where in activation space the state ends up. Representation is derived; future structure is prior. The evidence from Section 3 is consistent with this conjecture; establishing it in either direction would require constructing pairs of states and comparing their reachability distance with their activation-space distance, a comparison the framework now makes precisely actionable.

4.5. The Admissibility Field: First Introduction and Its Limitations

We are now in a position to introduce the central formal object of the paper. The reachability operator $R(a)$ is a rich object — a full probability distribution over trajectory space — and is difficult to work with directly. The natural first step is to summarize it by a scalar field. However, before doing so we must acknowledge an important subtlety.

The empirical observations described in Section 3 organize around *constrained continuation regimes*: copyright notices, patent headings, Q/A separators, import blocks. These are not characterized primarily by *validity* (all are perfectly valid text) but by *constraint density* — the degree to which the reachable future space is narrow, specific, and sharply bounded. A copyright notice and a generic prose sentence are equally "valid," yet the activation geometry treats them as categorically different because their continuation manifolds have radically different shapes.

This suggests that the natural scalar summary of $R(a)$ is not the probability of validity but something closer to the *geometry* of the reachable future set. Two natural candidates are the volume of the reachable set and its entropy:

$$A_{\mathrm{vol}}(x) = \mathrm{Vol}(R(x)), \qquad A_{\mathrm{ent}}(x) = H\big(R(x)\big), \tag{11}$$

where Vol is an appropriate measure on trajectory space and H is the differential entropy of the future distribution. Highly constrained states (patent headings, copyright notices) would have small A_{vol} and small A_{ent}; unconstrained generic states would have large values of both. These summaries are arguably closer to what the activation geometry is actually measuring than a probability of validity.

More fundamentally, the richer object underlying all of these scalar summaries is the *reachability bundle*:

$$\mathcal{R} = \big\{R(x)\big\}_{x \in \mathcal{M}}, \tag{12}$$

the collection of all reachability distributions, indexed by points on the activation manifold. The scalar admissibility field A is any scalar functional of $\mathcal{R}$ — one observable derived from a higher-order geometric object. The reachability bundle is the fundamental structure; the admissibility field is a projection. This observation, which we develop further in Section 11, suggests that the paper's central object may itself be a derived quantity of a deeper geometry.

For the purposes of organizing the present synthesis, we work with a general scalar admissibility field $A : \mathcal{M} \to \mathbb{R}_{\geq 0}$ without committing to a specific choice of functional form. The formal definition used throughout is the measure of the admissible trajectory set from each state, with the understanding that this can be

refined to volume, entropy, or other geometric quantities in future work.

Definition 4.4 (Admissibility Field)**.** Let $\mathcal{F}_{\text{valid}}$ be a measurable set of continuation trajectories considered admissible in a given context. The *admissibility field* is the scalar function

$$A : \mathcal{M} \longrightarrow \mathbb{R}_{\geq 0}, \qquad A(x) = \mu_x\big(R(x) \cap \mathcal{F}_{\text{valid}}\big), \tag{13}$$

where μ_x is the probability measure on trajectory space induced by the model's continuation dynamics starting from x. The choice of $\mathcal{F}_{\text{valid}}$ is context-dependent; empirically, the activation geometry suggests that the relevant criterion is structural constraint density (the shape of the future space) rather than semantic validity in any narrow sense.

This field will be the organizing object of the rest of the paper. We preview its roles: activation clusters will appear as level sets or basins of A (Section 5), ICA directions will appear as preferred gradient directions of A (Sections 6 and 7), and WMSD's velocity fields will appear as flows through A (Sections 8 and 9). The full synthesis, including the reachability bundle perspective, will be assembled in Section 11.

5. The Giant Component and Admissibility Topology

5.1. The Component Spectrum

The full component spectrum of Liu's graph G_θ contains roughly 10^6 activations. The dominant feature is a single giant connected component containing approximately 826,000 activations — about 82.6% of the total. The remaining 17.4% of activations are distributed across a large number of smaller components, ranging from mid-sized clusters of hundreds to thousands of activations down to isolated singletons.

A first reaction might be that the giant component is uninteresting — that it is simply the "general language" blob and the interesting structure is entirely in the smaller components. This reaction is understandable but misses something important. The structure of the *entire* spectrum, including the relative sizes of the giant component and the distribution of the smaller components, is itself informative. After removing the giant component, the residual component size distribution appears approximately power-law:

$$P(\text{component size} = k) \sim k^{-\alpha} \tag{14}$$

for some exponent $\alpha > 1$. Power-law distributions in graph component spectra are a signature of critical phenomena.

5.2. The Percolation Analogy

The emergence of a giant connected component in a network with a power-law residual spectrum is the hallmark of a percolation transition. In percolation theory [31], a random geometric graph on n points in $\mathbb{R}^d$ — where two points are connected whenever they are within distance r of each other — undergoes a phase transition as r varies. Below the critical threshold r_c, the graph consists of many small disconnected clusters. Above r_c, a single giant connected component emerges that contains a finite fraction of all vertices in the infinite-volume limit. At r_c itself, the component size distribution follows a power law.

The structure Liu observes is consistent with the graph being near or at such a critical threshold. The presence of a giant component containing the overwhelming majority of activations, combined with a power-law tail of small specialized components, suggests that the threshold θ is set in the vicinity of the percolation critical point for this particular activation distribution. This has a significant physical interpretation.

5.3. Admissibility Density and the Topological Decomposition

We can now interpret the component spectrum in terms of the admissibility field A. Define an *admissibility density* on $\mathcal{M}$ as follows: high-A regions are those where the continuation constraints are sufficiently universal that activations from many different contexts have overlapping reachability operators and thus cluster together. Low-A regions are those where the continuation is so highly constrained and specific that only activations from very similar specialized contexts have overlapping reachability operators.

The topological decomposition of the activation manifold is then:

$$\mathcal{M} = G \cup \bigcup_{i=1}^{\infty} \mathcal{I}_i, \tag{15}$$

where G is the giant component (the generic semantic ocean) and the $\mathcal{I}_i$ are the specialized admissibility islands. The giant component consists of states where A is smoothly varying and broadly connected: generic prose, general language, unrestricted continuation contexts. The islands consist of states where A changes rapidly over short distances because the continuation constraints shift abruptly from those of the generic language into a specialized mode.

This gives a precise relationship between the component structure and the admissibility field:

$$\text{component structure of } G_\theta \;\approx\; \text{level-set topology of } A(x). \tag{16}$$

The level sets $\{x : A(x) = c\}$ for different values of c trace out the boundaries between regions of different constraint density. The connected components of G_θ are approximations to the path-connected components of the superlevel sets $\{x : A(x) \geq c\}$ for an appropriate c.

5.4. Admissibility Islands as Linguistic Ecosystems

The specialized islands identified in Liu's analysis — patent language, legal citation structures, software licenses, LaTeX equation environments — are revealing. Each constitutes what we might call a *linguistic ecosystem*: a sub-register of text with unusually strong internal constraints and unusually low coupling to the surrounding activation space. The internal constraint density is high: given a state inside the patent-language island, the next state is almost certainly also inside that island. The external coupling is low: it is difficult to reach the patent-language island from generic prose contexts or from programming language contexts without passing through specific boundary tokens.

These properties define an isolated submanifold in activation space. The activations corresponding to patent text form a cluster not because patent text is semantically homogeneous — it is not, covering engineering, chemistry, biology, software, and countless other subjects — but because the *structural rules* governing patent text create a uniform admissibility landscape across all those subject matters. The level sets of A cut across subject-matter distinctions and group together anything that sits inside the same constraint field.

5.5. Connection to RSVP Phase Structure

The percolation picture also connects naturally to phase structure in field theories. In the RSVP framework [9], the admissibility field can undergo phase transitions as the parameter governing constraint density crosses a critical value. Near the critical point, the correlation length diverges and the component spectrum becomes scale-free — which is precisely the power-law behavior observed. The giant component is the "high-density phase" of the admissibility field; the isolated islands are "nucleated phases" corresponding to local maxima of the constraint potential embedded in a lower-density background.

5.6. Third Structural Observation

The third structural observation is: the component spectrum of the activation graph encodes the level-set topology of the admissibility field A. The giant component corresponds to the high-connectivity basin of A; the specialized islands correspond to isolated high-constraint regions where the admissibility landscape is steep and the crossing energy is high. The power-law distribution of component sizes is consistent with the graph being near a percolation critical point.

6. Constraint Fields and Natural Coordinates: A Bridge

6.1. Two Descriptions of the Same Geometry

At this point in the argument we have two empirical programs — Liu's activation graph and Liu & Han's ICA decomposition — and a single theoretical object A. It might appear that the graph-based and ICA-based descriptions are independent discoveries that require separate theoretical accounts. In this section we argue that they are in fact dual descriptions of the same geometric object, seen from two fundamentally different perspectives: one topological and one differential.

The key mathematical distinction is the following. A connected component of a graph is a *topological* object: it depends only on the qualitative question of whether a path exists between two vertices, not on the quantitative local geometry. A direction recovered by ICA is a *differential* object: it is a tangent direction to the manifold at a point, and its relevance depends on how the distribution varies locally. The graph asks which points belong together; ICA asks along which directions the distribution fractures.

6.2. Level Sets and Gradient Directions

Suppose the activation manifold $\mathcal{M}$ is embedded in $\mathbb{S}^{d-1}$ and the admissibility field $A : \mathcal{M} \to \mathbb{R}_{\geq 0}$ is a smooth scalar function on this manifold. The gradient $\nabla A(x)$ at a point $x \in \mathcal{M}$ is a tangent vector pointing in the direction of steepest ascent of A. The level sets $\{x : A(x) = c\}$ are the $(d-2)$-dimensional hypersurfaces perpendicular to ∇A.

Connected components of the graph G_θ are approximating the connected components of the superlevel sets of A. These are precisely the level-set topology. ICA directions, on the other hand, are seeking the directions in the tangent space along which the projection distribution is maximally non-Gaussian. We will argue in Section 7 that constraint selectivity implies heavy-tailed projections, which means that ICA directions are approximating the directions of largest variation in

A:

$$u_k \approx \text{directions of large variation in } \nabla A(x). \tag{17}$$

Graph components thus reveal the topology of A — which regions are connected, which are separated — while ICA directions reveal the differential structure of A — where it varies most rapidly and in which directions. They are complementary instruments for probing the same geometric object.

6.3. The Fundamental Geometric Triad

This duality positions us to state the central structural observation of the bridge section. The three empirical programs of this paper instantiate exactly the three foundational layers of modern differential geometry:

Topology asks what is connected, what is separated, what paths exist between points, how many holes a space has. This is the domain of Liu's activation graph: which activations belong to the same component, which components are isolated, what is the genus of the resulting graph structure.

Differential geometry asks what directions exist locally, what the curvature is, how the tangent spaces at different points are related. This is the domain of ICALens: which directions are preferred by the distribution, what is the statistical geometry of the local tangent space, which directions are most selective.

Dynamics asks how trajectories evolve through the space, what the vector fields are, what the attractors and repellers are. This is the domain of WMSD: what velocity field governs motion through possibility space, how that field is learned and transferred, what its stability properties are.

The correspondence is:

$$\text{Topology} \longrightarrow \text{Differential Structure} \longrightarrow \text{Dynamics} \tag{18}$$

$$\text{Liu's graph} \longrightarrow \text{ICALens directions} \longrightarrow \text{WMSD velocity fields} \tag{19}$$

and all three have the admissibility field A as their common underlying object:

$$\text{graph components} \approx \text{level sets of } A, \qquad \text{ICA directions} \approx \text{gradient directions of } A, \qquad \text{velocity} \tag{20}$$

This triad is not merely an aesthetic observation. Modern differential geometry is organized precisely in this order — topology first, differential structure second, dynamics third — because each layer presupposes and refines the previous one. Topology establishes what regions exist; differential geometry adds local quantitative structure to those regions; dynamics adds the temporal evolution that connects

points within and between regions. The fact that three independent empirical programs in mechanistic interpretability and world modeling map onto these three layers is strong circumstantial evidence that the mathematical structure of the activation manifold is genuinely geometric in a deep sense.

7. ICA as Natural Cleavage: Statistical Fractures in Activation Space

7.1. The ICALens Setup

The ICALens paper [23] takes a different entry point into the same geometry. Rather than constructing a graph over activations, it applies Independent Component Analysis (ICA) [17, 18] to the normalized and whitened activation matrix. The paper evaluates this approach across three model families: GPT-2 Small ($d = 768$), Gemma 2 2B ($d = 2304$), and Qwen 3.5 2B Base ($d = 2048$), fitting ICA independently at each residual-stream layer using one million token positions drawn from the Pile-10k dataset.

The preprocessing pipeline addresses a fundamental challenge: raw transformer activations are highly anisotropic, containing systematic outlier dimensions, rare massive activations, and attention-sink tokens that destabilize standard ICA fitting. The ICALens pipeline proceeds in three steps. First, each activation vector is row-normalized by its ℓ_2 norm:

$$r(x_i) = \frac{x_i}{\max(\|x_i\|_2, \varepsilon)}, \tag{21}$$

reducing the influence of activation-norm outliers before any further processing. Second, the row-normalized matrix is centered and whitened to remove second-order correlations. Third, a GPU-parallel FastICA algorithm is applied to the whitened matrix with a robust convergence criterion (the p95-LIM rule, which accepts a fit when 95% of components have stabilized even if a small tail remains difficult to converge) and an adaptive refit procedure that reduces the target component count for layers where full-rank convergence fails.

This engineering makes the difference: on GPT-2 Small with one million activations, the ICALens pipeline increases the number of accepted layers by 400% compared to a naïve FastICA implementation, and the resulting components are validated through human annotation, targeted prompt tests, and SAEBench evaluations.

The key interpretive claim of the paper is: *interpretable directions are statis-*

tically exceptional directions, and statistical exceptionality is measured by non-Gaussianity of the scalar projection distribution. The paper also establishes empirically that SAE decoder directions, despite being trained for sparse reconstruction rather than non-Gaussianity, already exhibit elevated kurtosis relative to random directions — suggesting that non-Gaussianity is a common statistical signature of learned feature directions, which ICA makes explicit as its direct optimization target.

7.2. Non-Gaussianity as Constraint Selectivity

Let $v \in \mathbb{S}^{d-1}$ be a candidate direction and let $z_i(v) = \left\langle v, \hat{h}_i \right\rangle$ be the scalar projection of the i-th normalized activation onto v. Define the excess kurtosis of this projection as

$$\kappa(v) = \frac{\frac{1}{n}\sum_{i=1}^{n}\big(z_i(v) - \bar{z}(v)\big)^4}{\left(\frac{1}{n}\sum_{i=1}^{n}\big(z_i(v) - \bar{z}(v)\big)^2\right)^2} - 3, \tag{22}$$

where $\bar{z}(v) = \frac{1}{n}\sum_i z_i(v)$ is the mean projection. For a Gaussian distribution, $\kappa(v) = 0$; heavy-tailed distributions have $\kappa(v) > 0$. FastICA maximizes $|\,\kappa(v)|$ (or a smooth approximation thereof) to find the most non-Gaussian projection directions.

Why should interpretable directions be non-Gaussian? The argument runs as follows. If a direction v is relevant to a specific constraint regime — say, the ICA direction for question-and-answer boundaries — then for the vast majority of tokens in the dataset, that direction will project to values near zero (since most tokens are not at Q/A boundaries). For the small fraction of tokens that are at Q/A boundaries, the projection will be large. The resulting distribution is quiet for most tokens and extreme for a few: precisely a heavy-tailed, high-kurtosis distribution.

By contrast, a direction that is relevant to generic semantic content would be activated moderately across a large fraction of tokens — different tokens activate it to different degrees, but not with the extreme sparsity of constraint-regime directions. Such a direction would have a projection distribution closer to Gaussian.

The chain of implication is therefore:

$$\text{constraint-regime selectivity} \Rightarrow \text{sparse activation} \Rightarrow \text{heavy-tailed projection} \Rightarrow \text{ICA recoverabili} \tag{23}$$

ICA recovers constraint-regime-selective directions not because it was designed with constraints in mind, but because the statistical footprint of constraint selectivity

— sparsity — produces the non-Gaussianity that ICA maximizes.

7.3. ICA Directions as Gradient Directions of the Admissibility Field

We can now connect the ICA directions to the differential structure of A. In the differential-geometric picture, the admissibility field A varies smoothly over the activation manifold $\mathcal{M}$. At any point $x \in \mathcal{M}$, the gradient $\nabla A(x)$ points in the direction of fastest increase of A. Directions along which A changes rapidly are directions where the constraint landscape is unstable: crossing a small distance in that direction moves from one constraint regime to a qualitatively different one.

ICA's non-Gaussianity criterion selects directions that are most selective between high-A and low-A states. For a direction u perpendicular to a level set of A, the projection $z_i(u)$ will be large for activations on one side of the level set and small for activations on the other side: this is precisely the sparse, bimodal structure that produces high kurtosis.

Therefore, ICA directions approximate the directions of large variation in ∇A at representative points:

$$u_k \;\approx\; \arg\max_{v \in \mathbb{S}^{d-1}} \mathbb{E}_{x \in \mathcal{M}} \big[\|\nabla_v A(x)\|^2 \big], \tag{24}$$

where $\nabla_v A(x) = \langle \nabla A(x), v \rangle$ is the directional derivative along v. The ICA decomposition is, in this light, approximating the eigendirections of the Hessian of the expected admissibility variation — the natural coordinate axes of the admissibility landscape.

7.4. Effective Receptive Field as Correlation Length

A significant technical contribution of [23] is the *effective receptive field* (ERF) diagnostic. ERF asks a precise question: how much left context is sufficient to recover the same signed component response at a target token? For a component j, an evidence set $\mathcal{E}_j$ is formed from examples whose absolute component score exceeds half the maximum score for that component. For each evidence example (x, t) in $\mathcal{E}_j$, suffixes of increasing length $x_{t-k+1:t}$ for $k = 1, 2, \ldots, K_{\max}$ are constructed and the component is recomputed. The sample-level ERF is the shortest suffix length k such that component j remains among the top-15 most active components and preserves its original sign:

$$\mathrm{erf}_j(x,t) = \min\{k \in \{1,\ldots,K_{\max}\} : j \in \mathrm{Top}_{15}\Big(\big\{|s_r(h_t^{(k)})|\big\}_r\Big) \wedge \mathrm{sign}(s_j(h_t^{(k)})) = \mathrm{sign}(s_j(h_t(x_{1:t})))\}. \tag{25}$$

The component-level ERF is the mean over the evidence set:

$$\mathrm{ERF}(j) = \frac{1}{|\mathcal{E}_j|} \sum_{(x,t)\in\mathcal{E}_j} \mathrm{erf}_j(x,t). \tag{26}$$

The paper uses $K_{\max} = 11$ and evaluates across all three model families. Components with small ERF are sensitive only to the current token or a short local window; components with large ERF depend on broader discourse context.

The paper reports a layer-wise shift across all three model families: components in shallow layers are dominated by token-local directions (surface constraints — tokenization, local syntax, character-level patterns), while components in deeper layers show more medium- and long-context directions. Importantly, the paper notes this is a gradual shift in mixture proportions rather than a sharp handoff, and that the largest-ERF components (those requiring more than the 11-token window) are most prevalent in *middle* layers — making middle layers a particularly useful target for inspecting context-dependent structure. The paper also finds a consistent negative correlation between ERF and excess kurtosis (Spearman correlations of -0.41 to -0.50 across models): high-kurtosis components tend to have small ERF, because they are activated by narrow recurring patterns whose triggering condition is visible at the target token, making them easier to annotate and explain.

This is precisely the local-to-global transition predicted by the topology-differential structure-dynamics triad. Shallow layers are where the topology of the activation manifold is established — which tokens are connected to which. Deeper layers are where the dynamics play out — where long-range constraint fields become activated and govern the trajectory of the text over extended sequences.

In field-theoretic language, ERF is a measure of the *correlation length* ξ_k of the k-th ICA component. The correlation length of a field measures the spatial (or here, contextual) range over which fluctuations in that field are correlated. A component with $\mathrm{ERF}(u_k) = 5$ tokens has $\xi_k \approx 5$: its activation is determined by local context and decays rapidly with context length. A component with $\mathrm{ERF}(u_k) = 100$ tokens has $\xi_k \approx 100$: its activation is shaped by the document structure over long ranges.

The layer-wise shift from small to large ERF then corresponds to the well-known renormalization-group picture of a field theory: short-range fluctuations are integrated out in early layers, leaving only the long-range modes that govern macro-scale behavior. This connection deserves more formal development than we can provide here; we flag it as one of the open questions in Section 11.

7.5. ICA versus SAE: Natural Coordinates versus Trained Dictionaries

The [23] paper positions ICA as a "first lens" before training sparse autoencoders (SAEs) [7, 4]. The comparison is instructive for our purposes.

A sparse autoencoder learns a dictionary $\{d_1, \ldots, d_m\} \subset \mathbb{R}^d$ with $m \gg d$ such that each activation can be expressed as a sparse combination: $h \approx \sum_{k \in S} c_k d_k$ for a small index set S. The dictionary is trained to optimize a reconstruction-plus-sparsity objective, and the resulting features are interpretable as learned semantic or functional categories. Crucially, the dictionary is an artifact of the training procedure; the directions d_k have no privileged status as natural axes of the activation manifold independent of the SAE training.

ICA, by contrast, recovers directions that are statistically privileged in the *original* activation distribution — directions along which the marginal distribution of projections is maximally non-Gaussian. These directions are not learned by an auxiliary model; they are properties of the activation distribution itself. In the language of differential geometry, ICA is finding the natural cleavage planes — the directions along which the manifold most naturally fractures — while SAEs are constructing a learned coordinate system that may or may not align with those natural directions.

The empirical comparison reported in [23] is more nuanced and more favorable to ICA than one might expect. On SAEBench sparse probing across the eight standard concept datasets, ICA is competitive with publicly released SAE dictionaries (Gemma Scope, Qwen Scope) and consistently outperforms both ITDA (a training-light sparse-coding alternative) and PCA. Compact Matryoshka SAE variants with smaller dictionary sizes also remain below ICA under equivalent component budgets, which means the ICA advantage is not simply a function of dictionary size. On SAEBench Targeted Probe Perturbation — an intervention-based disentanglement metric — ICA is strongest relative to public SAEs at small-to-medium intervention budgets, where a small number of ICA components captures enough class-relevant information to support selective interventions without requiring search through a large overcomplete dictionary.

The directional overlap analysis also reveals that ICA and SAE are complementary rather than redundant. Most ICA components have a nearest-SAE neighbor with moderate cosine similarity (roughly 0.3–0.6 in the median across layers and models) rather than near-perfect alignment, indicating partial but not complete agreement. Moreover, SAE features often activate on localized tokens (consistent with the sparsity pressure in SAE training), while ICA components frequently vary more smoothly across spans of related tokens — tracking, for example, the

financial interpretation of "bank" across deposit, check, withdraw, and cash in the same sentence. This difference in activation pattern is not accidental: it reflects the different objectives of the two methods. SAEs maximize sparse reconstruction; ICA maximizes non-Gaussianity. The former favors localized, event-like features; the latter favors contextual, constraint-tracking features. Both are measuring the activation manifold, but from different angular positions.

This suggests that the natural coordinate system recovered by ICA captures a significant fraction of the functionally important structure in activation space at much lower computational cost (no gradient-based dictionary training, no hyperparameter search over sparsity levels, no per-layer retraining). From the admissibility-field perspective, this is the expected result: ICA is finding the eigendirections of ∇A, which are the natural coordinate axes of the constraint landscape; SAEs are learning a trained dictionary that approximates those axes from data, with the additional capacity to resolve finer-grained distinctions within each natural cleavage plane.

7.6. Fourth Structural Observation

The fourth structural observation: activation space contains a latent coordinate system prior to any trained decoder. ICA recovers this coordinate system by finding statistically exceptional directions in the pre-existing activation distribution. These directions correspond to constraint-regime boundaries in the admissibility field A, and their effective receptive fields measure the correlation length of the corresponding constraint modes. The layer-wise shift from small to large ERF corresponds to the progressive integration of short-range into long-range constraint fields.

8. WMSD: Velocity Fields Over Possibility Space

8.1. The Teacher-Student Asymmetry

World Model Self-Distillation [30] addresses a different but structurally related problem. The setting is instruction-conditioned video world modeling: given an unlabeled scene image and a short task prompt, a model must generate a plausible video trajectory showing successful task execution. The paper evaluates on general tasks drawn from video-game environments and real-world scenes (the WorldTasks-Bench benchmark) and additionally transfers to robotic manipulation tasks on the DreamGen benchmark. The examples of tasks include cutting vegetables, stacking objects, navigating environments, and object interactions across first-

person, human-character, and robot agent settings. The training framework involves two agents with different information access:

The *Demonstrator* (teacher) receives rich conditioning: $c_D = (\mathcal{I}, \mathcal{D})$, where $\mathcal{I}$ is the initial observation (a scene image) and $\mathcal{D}$ is a detailed step-by-step solution description generated by a VLM showing exactly how the task is to be performed.

The *Executor* (student) receives impoverished conditioning: $c_E = (\mathcal{I}, \mathcal{T})$, where $\mathcal{T}$ is a short task prompt. The Executor sees the same initial image but receives no procedural guidance; it knows only the goal, not the path.

A concrete example drawn from the paper: for the task "cut the carrots," the Demonstrator receives a detailed solution description — "walk to the counter, pick up the knife, grasp the carrot, slice repeatedly, place the pieces in the bowl" — generated by a VLM given the initial scene image. The Executor receives only "cut the carrots" as the task prompt together with the same initial frame. The informational asymmetry is enormous, and yet the experimental results show the Executor eventually surpassing the Demonstrator under VLM-based evaluation.

8.2. What the Student Must Learn

The training objective for the Executor is not to imitate the Demonstrator's trajectories directly. Instead, the Executor's video world model is trained to generate plausible future trajectories from the current state given the goal, and a reinforcement learning signal from a vision-language model (VLM) provides feedback on whether the generated trajectories are successful.

The critical point is what the Executor must represent in order to perform well. It cannot memorize a trajectory τ^*, because it does not have access to the demonstration. It must instead learn the *region of trajectory space* that successfully achieves the goal — the set of trajectories that the Executor's generator could produce that would score positively under the VLM reward. This is not a single path through the workspace; it is a manifold of admissible paths.

In the language of the reachability framework, the Executor must learn an approximation to the set $R(x)$ of futures reachable from the current state that are compatible with the task goal. This is exactly the reachability operator from Section 4, now instantiated in the concrete setting of instruction-conditioned video generation.

8.3. The Velocity Field Formalism

The WMSD architecture models trajectories using conditional flow-matching video generators [30]. The latent video state $x_t \in \mathbb{R}^d$ evolves from a Gaussian base

distribution $x_0 \sim p_0$ at flow time $t = 0$ to the generated video latent $x_1 \sim p_1$ at $t = 1$. Concretely, the Executor generates a future trajectory by integrating a learned velocity field:

$$\frac{dx_t}{dt} = v_\theta(x_t, t \mid c_E), \tag{27}$$

where x_t is the state at time $t \in [0, 1]$, v_θ is the learned velocity field parameterized by θ, and c_E is the Executor's conditioning. Similarly, the Demonstrator's velocity field is

$$\frac{dy_t}{dt} = v_{\theta'}(y_t, t \mid c_D). \tag{28}$$

The fundamental object in this architecture is not a state representation but a vector field over state space. The "world model" is, in its mathematical essence, a velocity field $v(x, t)$ that governs how the state x evolves over time $t \in [0, 1]$.

This is a significant ontological choice. Classical planning models represent the world as a sequence of states connected by actions. The WMSD architecture represents the world as a continuous flow through a state manifold. In the former view, the primitive object is the state; in the latter, the primitive object is the velocity field. This is closely parallel to the ontological distinction in physics between particle mechanics (where the primitive object is the trajectory) and field theory (where the primitive object is the field itself). WMSD has, perhaps unwittingly, adopted the field-theoretic formulation.

8.4. Proposition 1: Local Agreement Controls Global Reachability

The technical heart of the WMSD paper is the stability theorem relating local velocity field agreement between Demonstrator and Executor to global trajectory divergence. The paper distinguishes two variants of distillation training that motivate the theorem. In *off-policy* distillation, the student velocity is matched to the teacher velocity on teacher-generated states: this is stable but constrains the student only on the teacher's trajectories, so errors compound during student rollouts. In *on-policy* distillation, the discrepancy is evaluated on student-generated states, addressing the distribution shift at the cost of a more complex gradient. It is the on-policy variant that motivates Proposition 1, and the empirical results show that on-policy self-distillation continues improving past the point where off-policy training saturates.

We reproduce the key derivation here and provide an interpretation that is not present in the original paper.

Theorem 8.1 (Velocity Field Stability; [30])**.** *Let v_θ and $v_{\theta'}$ be the Executor and Demonstrator velocity fields respectively, and suppose $v_{\theta'}$ is L-Lipschitz in its first*

argument. Let $x_0 = y_0$ be the shared initial state. Let $e_t = x_t - y_t$ be the trajectory error at time t. Suppose that

$$\int_0^1 \|v_\theta(x_t, t) - v_{\theta'}(x_t, t)\|^2 \, dt \leq \varepsilon^2. \tag{29}$$

Then the terminal Wasserstein-2 distance satisfies

$$W_2\big(p_\theta(x_1 \mid c_E),\ p_{\theta'}(x_1 \mid c_D)\big) \leq e^L \varepsilon. \tag{30}$$

Derivation. By the triangle inequality for vector norms applied to the trajectory error,

$$\frac{d}{dt}\|e_t\| \leq \left\| \frac{d}{dt}(x_t - y_t) \right\| = \|v_\theta(x_t, t) - v_{\theta'}(y_t, t)\|\,. \tag{31}$$

Adding and subtracting $v_{\theta'}(x_t, t)$:

$$\frac{d}{dt}\|e_t\| \leq \|v_\theta(x_t, t) - v_{\theta'}(x_t, t)\| + \|v_{\theta'}(x_t, t) - v_{\theta'}(y_t, t)\| \tag{32}$$

$$\leq \delta_t + L\,\|e_t\|\,, \tag{33}$$

where $\delta_t = \|v_\theta(x_t, t) - v_{\theta'}(x_t, t)\|$ is the local matching discrepancy and we have used the L-Lipschitz condition on $v_{\theta'}$. The Grönwall inequality applied to this differential inequality, with initial condition $\|e_0\| = 0$ (since $x_0 = y_0$), yields

$$\|e_1\| \leq e^L \int_0^1 \delta_t \, dt. \tag{34}$$

By the Cauchy-Schwarz inequality,

$$\int_0^1 \delta_t \, dt \leq \left(\int_0^1 \delta_t^2 \, dt \right)^{1/2} \leq \varepsilon, \tag{35}$$

where the last inequality uses assumption (29). Since the trajectory distributions are determined by the velocity fields, the bound on $\|e_1\|$ translates to the Wasserstein-2 bound (30) via the transport argument developed in the following subsection. □

8.5. From Trajectory Stability to Distributional Stability

The Grönwall argument above establishes a sample-wise bound: $\|x_1 - y_1\| \leq e^L \varepsilon$ when $x_0 = y_0$. The step to a Wasserstein-2 bound on the terminal distributions $p_\theta(x_1 \mid c_E)$ and $p_{\theta'}(x_1 \mid c_D)$ requires an explicit coupling argument. We provide it here.

The probability density path p_t for each flow is governed by the *continuity equation*:

$$\partial_t p_t + \nabla \cdot (p_t v_t) = 0, \tag{36}$$

which links the evolution of probability mass to the velocity field v_t. The Wasserstein-2 distance between two distributions μ and ν on $\mathbb{R}^d$ is defined via optimal transport:

$$W_2(\mu,\nu)^2 = \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|^2 \, d\gamma(x,y), \tag{37}$$

where $\Pi(\mu,\nu)$ is the set of couplings with marginals μ and ν.

We construct a specific coupling. Let $x_0 \sim p_0$ be the shared base noise. Under the Executor's flow with velocity v_θ, the point x_0 maps to x_1^θ. Under the Demonstrator's flow with velocity $v_{\theta'}$, the same x_0 maps to $x_1^{\theta'}$. The joint distribution of $(x_1^\theta, x_1^{\theta'})$ under this shared-noise coupling is a valid element of $\Pi(p_\theta(x_1 \mid c_E), p_{\theta'}(x_1 \mid c_D))$.

Using this coupling,

$$W_2\big(p_\theta(x_1 \mid c_E), p_{\theta'}(x_1 \mid c_D)\big)^2 \le \mathbb{E}_{x_0 \sim p_0}\big[\big\|x_1^\theta - x_1^{\theta'}\big\|^2\big] \tag{38}$$

$$= \mathbb{E}_{x_0 \sim p_0}\big[\|e_1\|^2\big]. \tag{39}$$

The Grönwall bound gives $\|e_1\| \le e^L \varepsilon$ for each realization of x_0, so

$$W_2\big(p_\theta(x_1 \mid c_E), p_{\theta'}(x_1 \mid c_D)\big) \le \sqrt{\mathbb{E}[\|e_1\|^2]} \le e^L \varepsilon, \tag{40}$$

where the last inequality uses the deterministic bound on $\|e_1\|$.

The continuity equation (36) plays a central structural role here: it is the mechanism by which local velocity field agreement propagates to global distributional agreement. Matching velocity fields means matching the vector fields that transport probability mass; the Wasserstein distance between the terminal distributions is therefore bounded by the integrated mismatch in those transports. This connection between the continuity equation and Wasserstein stability is precisely the bridge between the differential and dynamical layers of the geometric triad established in Section 6.

The standard reading of this theorem is as a technical stability result: if you train the Executor to match the Demonstrator's velocity field pointwise, the Executor's terminal distribution will be close to the Demonstrator's.

Our reading is different and, we argue, more fundamental. Condition (29) says that the Executor's velocity field agrees with the Demonstrator's on the Executor's

own trajectories. This is a condition on *local admissibility agreement*: at every point the Executor visits, the direction of evolution it chooses must agree with the direction the Demonstrator would choose. Result (30) says that this local agreement controls global trajectory distributions. In our language:

> **Local Admissibility Agreement Controls Global Reachability.** If the Executor matches the Demonstrator's velocity field on its own trajectories (local condition), then the Executor's terminal state distribution is close to the Demonstrator's (global consequence). Local velocity agreement induces global trajectory agreement.

This is precisely the statement one would expect in a framework where the fundamental object is the admissibility field A rather than individual states. The velocity field $v_\theta(x, t)$ encodes, at each state x, the locally admissible direction of motion. Matching velocity fields means matching the local admissibility structure; the global result follows by integration.

8.6. Generation-Verification Asymmetry as Admissibility Filtering

A recurring theme in the WMSD paper is the *generation-verification asymmetry*: it is much harder to generate a solution trajectory than to verify whether a proposed trajectory achieves the goal. This asymmetry is pervasive in computational problems [26] — SAT solving, theorem proving, protein folding — and the WMSD architecture exploits it by separating the generation role (the Executor's world model, which proposes trajectories) from the verification role (the VLM, which assesses whether proposed trajectories are successful).

In the admissibility framework, this is an instantiation of admissibility filtering. The Executor proposes a trajectory τ; the VLM checks whether τ belongs to the admissible set $\mathcal{F}_{\text{valid}}$. In the actual WMSD implementation, the VLM (Qwen3.5-27B in the primary experiments) produces a binary judgment on task completion, and the task reward is defined as the log-probability difference:

$$r_{\text{task}}(\tau) = \log p_{\text{VLM}}(\texttt{yes} \mid \tau) - \log p_{\text{VLM}}(\texttt{no} \mid \tau), \tag{41}$$

which encodes both the predicted outcome and the model's uncertainty. The paper also introduces a consistency reward that penalizes violations of physical plausibility and temporal coherence, providing a safeguard against reward hacking (unrealistic object appearances, implausible motion). The VLM is therefore not acting as a hard admissibility oracle but as a *soft* admissibility signal — a noisy but useful measure of how deeply inside the admissible set $\mathcal{F}_{\text{valid}}$ the proposed

trajectory lies. The combination of distillation regularization and soft VLM reward gives the system a well-defined interior to explore rather than a sharp boundary to exploit.

8.7. Fifth Structural Observation

The fifth structural observation: the WMSD architecture instantiates the dynamics layer of the topology-differential structure-dynamics triad. Its fundamental primitive is a velocity field $v(x, t)$ over possibility space, not a state representation. The key stability theorem is a statement about local admissibility agreement controlling global reachability. The generation-verification asymmetry is a computational instance of admissibility filtering. WMSD is not a world model in the classical state-representation sense; it approximates an admissibility field.

9. Distillation as Admissibility-Preserving Projection

9.1. The Compression Question

The WMSD result that the Executor eventually surpasses the Demonstrator's performance appears paradoxical. In ordinary distillation settings, the student is expected to inherit the teacher's performance ceiling when trained only to imitate the teacher directly, since the student has access to strictly less information at inference time. How can the Executor, who receives only the task prompt $\mathcal{T}$ while the Demonstrator receives the full solution description $\mathcal{D}$, come to perform better than the Demonstrator?

The resolution lies in recognizing that the Demonstrator is not the objective. The Demonstrator is a *prior*: it biases the Executor's search toward the manifold of successful trajectories, but the ultimate arbiter of success is the VLM reward r_{task}. Once the Executor has used the Demonstrator's signal to initialize near the relevant region of trajectory space, the reinforcement learning component can explore within that region and find trajectories that are strictly better than any the Demonstrator can produce.

The deep question is: what structural information is preserved when the Demonstrator's rich signal $(\mathcal{I}, \mathcal{D})$ is compressed to the Executor's sparse signal $(\mathcal{I}, \mathcal{T})$? And why is enough structure preserved for the Executor to succeed?

9.2. The Task Sufficiency Projection Principle

Define the compression map:

$$\pi : (\mathcal{I}, \mathcal{D}) \;\mapsto\; (\mathcal{I}, \mathcal{T}). \tag{42}$$

This map discards the detailed solution $\mathcal{D}$ and retains only the task description $\mathcal{T}$. It is a projection from a high-information to a low-information representation. Let $R(\mathcal{I}, \mathcal{D})$ and $R(\mathcal{I}, \mathcal{T})$ denote the reachability operators associated with the rich and sparse conditioning signals respectively — the distributions over future trajectories accessible under each conditioning.

Task Sufficiency Projection Principle. The compression π is *admissibility-preserving* when

$$D\big(R(\mathcal{I}, \mathcal{D}),\; R(\mathcal{I}, \mathcal{T})\big) \leq \varepsilon \tag{43}$$

for sufficiently small ε. The empirical success of WMSD demonstrates that π is approximately admissibility-preserving for the tasks studied: despite the enormous informational asymmetry, the reachability structure is approximately preserved.

This principle has a precise interpretation in terms of the admissibility field. The projection π is admissibility-preserving if the level sets of A in the Demonstrator's conditioning space are approximately aligned with the level sets of A in the Executor's conditioning space. The detailed solution $\mathcal{D}$ specifies a particular trajectory through A's landscape; the task specification $\mathcal{T}$ specifies only the target region of that landscape. The projection preserves which region is the target, even though it discards the specific path.

9.3. The Role of the Reward in Reshaping the Geometry

The WMSD reward decomposes as:

$$R(\tau) = \lambda_{\text{task}}\, r_{\text{task}}(\tau) + \lambda_{\text{distill}}\, r_{\text{distill}}(\tau). \tag{44}$$

The distillation reward r_{distill} measures how closely the Executor's trajectory matches the Demonstrator's distribution. The task reward r_{task} measures whether the trajectory achieves the goal. Their combination has the geometric interpreta-

tion:

$$\text{local prior (Demonstrator)+global direction (reward)} \Rightarrow \text{admissibility-guided trajectory search.} \tag{45}$$

The distillation term provides local geometry: it ensures that the Executor stays near the manifold of trajectories the Demonstrator considers plausible, preventing the RL component from wandering into degenerate regions of state space. The reward term provides global direction: it reshapes the topology of the admissibility field toward goal satisfaction, expanding the basin of attraction around successful trajectories.

Together, these two terms define a constraint field that the Executor learns to navigate. The Executor is not copying trajectories; it is learning the geometry of admissible trajectory space as jointly defined by the Demonstrator's prior and the reward's reshaping. This is why it can surpass the teacher: the teacher's trajectory was one point in the admissible manifold; the Executor, having learned the geometry of the entire manifold, can find better points.

9.4. Detailed Instructions as Projection Artifacts

A key conceptual conclusion of the Task Sufficiency Projection Principle is that the detailed solution $\mathcal{D}$ is, in a precise sense, a *projection artifact* of the task representation. The admissibility structure — which trajectories lead to goal satisfaction — is fully captured by the task specification $\mathcal{T}$ together with the task reward, without requiring the detailed procedural walkthrough. The Demonstrator's solution provides a useful initialization and regularization signal, but the information that is causally necessary for task competence is already present in the task itself.

This is a non-trivial claim. It might have been the case that the detailed procedural steps contain information about the task that is not recoverable from the task description alone — that certain subtleties of execution are specified by the Demonstrator that the Executor could not independently discover. The empirical result says otherwise: the constraint structure of the task is largely determined by the task specification, and the Executor can learn that structure through exploratory reinforcement learning guided by the reward signal. Procedure is a projection artifact; constraint is invariant.

This is a direct parallel to the claim in the reachability framework [12] that much of the representational content of a system can be derived from its future-reachability structure, without requiring an explicit specification of all the intermediate steps.

10. Frozen Processes and the Ontological Inversion

10.1. A Recurring Pattern

We have now examined three independent empirical programs in detail. In each case, we have identified the structural observation that the program is making, connected it to the admissibility field A, and shown how the connection illuminates both the empirical finding and the formal theory. But there is a meta-level observation that has been implicit throughout and that deserves explicit articulation: all three programs, entirely independently, have performed the same ontological inversion.

By *ontological inversion* we mean the shift from a representation-primary account to a future-structure-primary account. In the traditional view of language models and intelligent systems, the basic ontological category is the *object*: the model represents things — tokens, concepts, facts, states of the world — and behavior is derived from those representations. In the picture that emerges from the three programs we have studied, the basic ontological category is the *transformation*: the model represents possibilities, constraints, and continuation dynamics, and objects appear as derived categories that inherit their identity from the transformational context in which they occur.

10.2. The Empirical Inversion Table

This inversion is not a metaphysical claim we are imposing on the data; it is a reading of what the data actually shows. The following summary captures the contrast:

The traditional view expects *concepts* as the basic units of representational organization: the network should have representations of "dog," "democracy," "prime number," and so forth. What the empirical programs find instead are *constraint regimes*: the network organizes around Q/A boundaries, patent headings, copyright notices — not around the concepts that the text happens to discuss.

The traditional view expects *features* as the basic units of activation structure: individual neurons or linear directions should correspond to identifiable properties of the input. What ICALens finds instead are *continuation modes*: directions that are activated specifically when the text enters a particular structural regime, not when a particular concept is instantiated.

The traditional view expects *states* as the basic units of world modeling: the system should represent the current configuration of the environment. What WMSD finds instead are *trajectories*: the system is organized around the space of future evolution, and states are points on trajectories rather than the fundamental

objects.

The traditional view expects *objects* as the basic ontological category: things with properties that persist through time. What all three programs find are *transformations*: structured modes of change that define the identity of the context they apply to.

10.3. Nouns as Frozen Verbs: A Mechanistic Reading

The thesis that nouns are "frozen verbs" — that object-concepts are derived from process-concepts by a kind of temporal averaging or crystallization — has a long history in process philosophy [19, 33]. In the context of the present paper, this thesis acquires a precise mechanistic reading.

Consider again the Rank-2 cluster: 276 instances of the colon ":" after "Q." The colon is, in standard linguistic analysis, a punctuation symbol with a lexical identity. But the activation analysis reveals that the network is not primarily representing "colon-as-symbol"; it is representing "colon-as-entrance-to-Q/A-regime" — a processual state defined by the transformation it initiates rather than by the object it labels.

Similarly, the patent-heading cluster is not primarily representing "heading-as-noun-phrase"; it is representing "heading-as-entry-into-patent- discourse-mode" — a structural transformation that reorganizes the admissibility landscape. The copyright cluster is not representing "copyright notice as a type of legal statement"; it is representing the transformation into copyright-language mode.

In each case, the noun-like entity (the lexical token, the surface syntactic structure) is secondary to the verb-like entity (the transformation mode, the constraint regime transition) that the network's geometry is actually tracking. The activation manifold is not partitioned into things; it is partitioned into families of admissible transformations.

10.4. The Inversion as Empirical Consequence

The crucial point is that this ontological inversion is not arrived at by philosophical argument. It is not that we have decided on a priori grounds that process-primary ontology is more elegant or more fundamental. It is that all three empirical programs — the activation graph, the ICA decomposition, and the world model distillation — have independently arrived at representations that are organized around transformations rather than objects. The inversion is forced by the data.

This gives us a stronger claim than process philosophy has previously been able to make. The argument is no longer: "conceptually, it is more natural to

think of nouns as derived from verbs." The argument is: "the internal geometry of large-scale neural language models and world models is, as an empirical matter, organized primarily around transformational modes rather than object categories. The traditional noun-primary ontology is not merely philosophically inelegant; it is descriptively inaccurate at the level of the network's actual computational structure."

11. The Admissibility Field Hypothesis

11.1. A Note on Ontological Status

Before assembling the full formal synthesis, we pause for an important qualification. The admissibility field $A : \mathcal{M} \to \mathbb{R}_{\geq 0}$ has been introduced and used throughout this paper as a theoretical organizing object. This usage risks a misreading that we want to explicitly forestall.

A is *not* assumed to exist as a primitive scalar quantity that can be directly measured from the model's weights or activations. Rather, it is introduced as the minimal latent object whose topology, gradients, and induced flows would jointly account for the observed clustering, ICA directions, and velocity-field dynamics. The hypothesis succeeds insofar as it generates testable predictions unavailable to purely representational accounts. If tomorrow the evidence were better explained by a tensor field, a bundle, or a sheaf, those objects would take A's place; the structural correspondence between the three empirical programs and the three layers of geometry would remain.

The admissibility field is the simplest scalar projection of an underlying geometric structure that we now make explicit.

11.2. The Reachability Bundle and the Reachability Metric

The most fundamental object in the framework is not the scalar admissibility field but the *reachability bundle*

$$\mathcal{R} = \big\{R(x)\big\}_{x \in \mathcal{M}}, \tag{46}$$

the collection of all reachability operators, one for each point on the activation manifold. The admissibility field A is one scalar functional of this bundle: $A(x) = F(R(x))$ for some function F (measure of admissible mass, volume, entropy, etc.). Different choices of F yield different scalar fields, all of them shadows of the same underlying object $\mathcal{R}$.

This reordering is significant. The paper's argument has been:

$$\mathcal{R} \longrightarrow A \longrightarrow \text{clusters, ICA directions, velocity fields.} \tag{47}$$

But the deeper claim, more consistent with the prior work in this series, is that $\mathcal{R}$ is the fundamental object and A is merely one observable:

$$\mathcal{R} \longrightarrow \text{topology of clusters} \quad \text{and} \quad \mathcal{R} \longrightarrow \text{differential structure of ICA} \quad \text{and} \quad \mathcal{R} \longrightarrow \text{dynamics} \tag{48}$$

with A playing the role of a useful coordinate system on $\mathcal{R}$ rather than the fundamental object itself.

The reachability bundle also suggests a natural Riemannian structure on $\mathcal{M}$. Define the *reachability volume* at x as

$$V(x) = \mathrm{Vol}\big(R(x)\big) \tag{49}$$

for some measure on trajectory space. Then define a metric tensor:

$$g_{ij}(x) = \frac{\partial^2}{\partial x_i \partial x_j} \log V(x), \tag{50}$$

the Hessian of the log-reachability-volume. This is the *reachability metric*: a natural Riemannian metric on activation space induced by the geometry of reachable futures rather than by any representational distance. Under this metric:

Activation clusters become regions of similar reachability volume — level sets and basins of $V(x)$ rather than of an externally defined validity measure.

ICA directions become principal curvature directions of the reachability metric — the eigenvectors of g_{ij} at representative points, pointing along the directions of most rapid variation in V.

WMSD velocity fields become geodesic flows or gradient flows through the reachability metric, following the natural geometry of the reachability landscape rather than an imposed objective.

This formulation tightens the mathematical backbone considerably: the activation manifold becomes a Riemannian manifold with a canonical metric derived entirely from the model's learned continuation structure, and all three empirical programs become geometric measurements of that manifold.

11.3. The Geometric Interpretability Correspondence

We now state formally the structural correspondence that has been the central architectural achievement of the paper. The following is not a theorem in the

strict sense — its precise formulation would require additional assumptions about smoothness and the form of $V(x)$ — but it is the central conceptual claim that all subsequent arguments elaborate.

Geometric Interpretability Correspondence. Let $\mathcal{M}$ be the normalized activation manifold and $\mathcal{R}$ the reachability bundle. The three major families of interpretability measurement correspond to the three foundational layers of differential geometry applied to $\mathcal{R}$:

$$\begin{aligned} \text{Topological probes (Liu's graph)} &\longleftrightarrow \text{level-set structure of } \mathcal{R} \\ \text{Differential probes (ICALens)} &\longleftrightarrow \text{gradient and curvature structure of } \mathcal{R} \\ \text{Dynamical probes (WMSD)} &\longleftrightarrow \text{flow and transport structure of } \mathcal{R} \end{aligned} \tag{51}$$

These three families are measuring the same underlying geometric object — the reachability bundle — from three different mathematical perspectives: topological (which regions are connected), differential (which directions are preferred), and dynamical (how states evolve). The three layers are logically ordered: topology precedes differential structure, which precedes dynamics.

This correspondence is the most original conceptual contribution of the paper. It survives even if the specific scalar field A is replaced by a better object. As long as the activation manifold has a reachability bundle, the three measurement families will correspond to the three layers of geometry applied to that bundle. The correspondence is structural, not dependent on any particular choice of scalar summary.

11.4. Formal Definition of the Admissibility Field

We now give the most precise version of the scalar admissibility field, with the understanding that this is one projection of the deeper reachability bundle $\mathcal{R}$.

Definition 11.1 (Admissibility Field, formal version)**.** Let $\mathcal{M} \subset \mathbb{S}^{d-1}$ be the normalized activation manifold, and let Ω be the space of all future activation trajectories $\tau = (x_{t_0}, x_{t_1}, \ldots) \subset \mathcal{M}$ beginning from various initial states. Let $\mathcal{F}_{\text{valid}} \subset \Omega$ be a measurable set of *admissible* trajectories. Let μ_x be the probability measure on Ω induced by the model's continuation dynamics starting from state x. The *admissibility field* is

$$A : \mathcal{M} \longrightarrow \mathbb{R}_{\geq 0}, \qquad A(x) = \mu_x\big(\mathcal{F}_{\text{valid}}\big) = \int_{\Omega} \mathbf{1}[\tau \in \mathcal{F}_{\text{valid}}]\, d\mu_x(\tau). \tag{52}$$

Equivalently, $A(x)$ is the probability, under the model's continuation dynamics starting from x, that the resulting trajectory is admissible. The admissibility field is one scalar observable derived from the reachability bundle; the reachability volume $V(x) = \mathrm{Vol}(R(x))$ and the reachability entropy $H(R(x))$ are alternative observables with arguably closer correspondence to the empirical constraint-density observations.

11.5. The Four Measurements

With the formal structure in hand, we can now state the claim that the three empirical programs are different measurements of $\mathcal{R}$ (or equivalently, of A as its scalar proxy):

$$\begin{aligned} \text{Activation clusters} &\approx \text{path-connected components of superlevel sets } \{x : A(x) \geq c\} \\ \text{ICA directions} &\approx \text{eigendirections of } g_{ij}(x) \text{ (or of } \nabla^2 A) \\ \text{Effective Receptive Field} &\approx \text{correlation length of } A \\ \text{WMSD velocity fields} &\approx \text{gradient flows through } V(x). \end{aligned} \tag{53}$$

11.6. The Synthesis Progression

The argument of this paper has proceeded in five steps. We now state them explicitly as a logical sequence:

Step 1. Normalization is a canonical projection $\pi_{\mathbb{S}}$ from $\mathbb{R}^d \setminus \{0\}$ to $\mathbb{S}^{d-1}$, collapsing the scale fiber $\mathbb{R}_{>0}$ and aligning the representation with the computationally relevant angular geometry. This projection reveals latent geometric structure.

Step 2. The revealed structure clusters not around semantic topics but around continuation constraints. Activations cluster together when they occur at the entrance to the same constrained continuation regime, not when they refer to the same objects or concepts.

Step 3. ICA finds statistically exceptional directions in the pre-existing activation distribution. These directions correspond to constraint-regime boundaries and approximate the curvature directions of the reachability metric. They are natural cleavage planes of the activation manifold.

Step 4. WMSD's learned velocity fields represent the dynamics of reachability-guided trajectory evolution. The fundamental stability result (Theorem 8.1) says that local velocity agreement — formalized through the continuity equation and shared-noise coupling — controls global distributional agreement.

Step 5. All four observations — clustering, ICA directions, ERF, velocity fields — are unified by the reachability bundle $\mathcal{R}$, of which A is one scalar projection. The admissibility field is the simplest geometric entity capable of organizing this collection of observations; the reachability bundle is the underlying geometric object of which it is an observable.

$$\boxed{\text{Features are emergent coordinates of admissibility topology.}} \tag{54}$$

11.7. The ERF Green's Function Conjecture

The open question of formalizing ERF as a correlation length can be given a more precise form as a conjectural field equation. Suppose the propagation of constraint intensity along the context axis s behaves like a massive scalar field satisfying a screened Poisson equation:

$$\left(\nabla_s^2 - m_k^2\right)\Phi_k(s) = J_k(s), \tag{55}$$

where Φ_k is the activation of ICA component k as a function of context position s, J_k is the source term (the local constraint trigger), and m_k is an effective mass. The Green's function for equation (55) in one dimension decays as

$$G_k(s, s_0) \sim e^{-m_k|s-s_0|}, \tag{56}$$

so the characteristic decay length is $\xi_k = 1/m_k$. We then conjecture:

$$\mathrm{ERF}(k) \;\approx\; \xi_k = \frac{1}{m_k}. \tag{57}$$

Under this conjecture, the empirical findings acquire a direct field-theoretic interpretation. High-kurtosis components with small ERF correspond to large effective mass $m_k \gg 0$: the constraint is strongly screened and decays rapidly over context. Long-context components with large ERF correspond to $m_k \approx 0$: the constraint behaves like a massless field and propagates over long ranges. The layer-wise shift from small to large ERF components corresponds, in renormalization group language, to the progressive appearance of near-massless modes as short-range fluctuations are integrated out in deeper layers.

This conjecture connects ERF to the correlation length of the admissibility field in a precise way and makes quantitative predictions: fitting the screened Poisson model to empirical ERF data would yield component-specific effective masses m_k whose distribution across layers and kurtosis values constitutes a testable

prediction.

11.8. Open Formal Questions

Beyond the Green's function conjecture, the admissibility field hypothesis raises several additional formal questions.

The first is the *computability* of A and $V(x)$. Definition 11.1 requires integration over the space of all future trajectories, which is exponentially large. Efficient estimation from finite samples and formal approximation guarantees are needed.

The second is the *renormalization interpretation* of the component spectrum. The power-law distribution of component sizes suggests proximity to a percolation critical point. Can this be formalized as a renormalization-group fixed point of the reachability bundle? Is there a natural coarse-graining operation on $\mathcal{M}$ under which $\mathcal{R}$ is approximately self-similar?

The third is the *curvature interpretation* of Theorem 8.1. The stability bound involves the Lipschitz constant L of the teacher velocity field. Can this bound be tightened using the sectional curvature of the reachability metric g_{ij}, replacing the global constant L with a local geometric invariant?

The fourth is the *reachability compression theorem*. The RDR conjecture claims that similar futures induce geometric proximity. A rigorous version would derive a bound of the form:

$$D\big(R(x), R(y)\big) \leq \varepsilon \quad \Rightarrow \quad \left\|\hat{h}_x - \hat{h}_y\right\| \leq f(\varepsilon) \tag{58}$$

from predictive sufficiency arguments about the training objective, without assuming it empirically. Such a theorem would make the RDR conjecture a consequence of the statistical structure of next-token prediction training, and would provide the foundational theorem from which all other results in the paper could be derived.

12. Predictions and Falsification Criteria

A theoretical framework earns its place not only by organizing known observations but by generating predictions that distinguish it from alternatives. The admissibility field hypothesis and the RDR conjecture make several empirical predictions that could in principle be tested with existing interpretability infrastructure.

Prediction 1: Reachability predicts activation neighborhoods better than semantic similarity. If the RDR conjecture is correct, then the reachability distance $D(R(a_1), R(a_2))$ should predict activation-space proximity $\left\|\hat{h}_{a_1} - \hat{h}_{a_2}\right\|$ better than any semantic similarity measure (topic similarity, distributional word

vectors, etc.). A falsification: systematic evidence that semantic similarity outperforms reachability distance as a predictor of activation-space proximity would challenge the RDR conjecture directly.

Prediction 2: ICA directions with highest kurtosis correspond to strongest future-space partition boundaries. If ICA directions approximate eigendirections of the reachability metric g_{ij}, then the most non-Gaussian components should correspond to the steepest boundaries between distinct continuation regimes. A falsification: high-kurtosis ICA directions found to correspond to arbitrary lexical distinctions rather than continuation-regime transitions would break the connection between non-Gaussianity and admissibility gradient structure.

Prediction 3: Distillation success correlates with reachability structure preservation. The Task Sufficiency Projection Principle predicts that distillation succeeds when $D(R(\mathcal{I}, \mathcal{D}), R(\mathcal{I}, \mathcal{T})) \leq \varepsilon$. A testable consequence: distillation should fail or degrade systematically when the task specification $\mathcal{T}$ loses information about the shape of the future trajectory manifold (e.g. when the task prompt is severely ambiguous) even if the representation of the goal state is preserved. A falsification: successful distillation in cases where the reachability structure is demonstrably not preserved.

Prediction 4: The ERF–kurtosis outliers are the most interpretively important components. The observed correlation between ERF and excess kurtosis ($\rho \approx -0.5$) is moderate, not decisive. This means there exist components with large ERF and large kurtosis simultaneously. These are components that depend on long-range context but activate only rarely and strongly. The admissibility framework predicts that these outliers correspond to the most important structural boundaries in the model: discourse-mode transitions, document-level structural boundaries, latent task switches — precisely the places where the admissibility landscape changes most sharply over long contextual ranges. A systematic annotation of these outlier components constitutes a targeted test of this prediction.

Prediction 5: Activation clusters predict continuation similarity better than semantic similarity. Liu's clusters are organized around constrained continuation regimes rather than semantic topics. A direct test: for any two tokens in the same cluster, their empirically observed continuation distributions should be more similar than their semantic similarity would predict. Conversely, two semantically similar tokens from different clusters should have more divergent continuation distributions than semantic similarity would suggest. This prediction can be tested directly using the activation graph from [22] paired with continuation-distribution measurements.

Once the framework generates predictions of this kind, the admissibility field

stops being merely an interpretation and becomes a research program. Each prediction constitutes a constraint on the reachability bundle $\mathcal{R}$ and each falsification would require a structural revision of the framework.

13. Conclusion

13.1. Summary of the Argument

We began with three empirical results that appeared, on their surfaces, to be about entirely different things: the geometry of activation vectors on a hypersphere, the statistical structure of independent component directions, and the distillation of step-by-step solutions into high-level task performance. We have argued that these three results are different measurements of a single underlying geometric object — the reachability bundle $\mathcal{R}$ over the activation manifold, summarized for present purposes by the scalar admissibility field $A : \mathcal{M} \to \mathbb{R}_{\geq 0}$ — and that recognizing this common origin explains the structural similarities among them and opens several new directions for both theory and practice.

The central conceptual claim is the Reachability Determines Representation conjecture: the organizational structure of the activation manifold is shaped by future-reachability proximity classes rather than by present-state feature categories. We have been careful throughout to distinguish what the evidence directly shows (geometrically clustered states share similar continuation structures) from what the conjecture asserts in the causal direction (similar futures induce geometric proximity). Both directions are plausible given the evidence; only the latter requires empirical verification beyond what the papers analyzed here provide.

The central formal contribution is the Geometric Interpretability Correspondence: the observation that the three empirical programs instantiate the three foundational layers of differential geometry — topology (Liu's graph), differential structure (ICALens directions), dynamics (WMSD velocity fields) — applied to the same underlying geometric object. This correspondence is structurally robust: it holds even if the specific scalar field A is replaced by a more refined object such as the reachability volume $V(x)$ or the full reachability bundle $\mathcal{R}$.

A secondary formal contribution is the clarification that the ε-future proximity relation $\approx_R^\varepsilon$ is a uniform tolerance structure rather than an equivalence relation for $\varepsilon > 0$, and the addition of an explicit continuity-equation derivation bridging the sample-wise Grönwall bound to the Wasserstein-2 distributional stability result of Theorem 8.1.

13.2. Implications for Interpretability

If the admissibility field hypothesis is correct, it has significant implications for the agenda of mechanistic interpretability. The standard agenda assumes that the natural decomposition of activation space is *featural*: the goal is to identify a set of directions or sparse features such that each activation can be expressed as a combination of interpretable feature-values. The dominant tools — sparse autoencoders, circuit analysis, feature visualization — are all designed to find and analyze these featural components.

Our argument suggests that this featural decomposition, while useful, may be an *emergent coordinate system* over a more fundamental topological structure. Features are not the primitive objects; they are the natural coordinates of the admissibility landscape. The primary structure is the topology of the level sets of A — which regions are connected, which are isolated, which are on the boundary between regimes — and features appear as the differential structure overlaid on this topology.

This suggests a different methodological priority: understand the topology of A first, then derive the feature coordinates from the topology. Liu's activation graph is already doing this: it is recovering the topology directly, without going through features. ICALens is recovering the differential structure, one level more refined than the topology but still more fundamental than a trained feature dictionary. A fully topological interpretability program would begin with the level-set structure of A and derive features as its natural coordinate system.

13.3. Implications for Alignment

The alignment implications are correspondingly reframed. Standard approaches to alignment focus on specifying rewards, preferences, and values, then training systems to optimize for them. This is the representational paradigm: we specify what we want (a representation of desired behavior) and train the system to match it.

If the internal geometry of intelligent systems is organized around admissibility fields rather than reward functions, the natural alignment target is not a reward specification but a *geometry of admissible futures*. We do not want to tell the system what is good; we want to shape the constraint field within which the system moves so that the admissible future set coincides with the futures we prefer. This is a geometric problem, not an optimization problem.

More precisely: alignment may depend less on specifying rewards and more on shaping the geometry of admissible futures. The technical challenge is to

understand A well enough to know what changes to training, architecture, or data would deform its level sets in the desired directions.

13.4. The Broader Convergence

A final word on the significance of the convergence itself. In the history of science, convergences of this kind — where independent empirical programs arrive at the same formal structure from different directions — are often signals that the formal structure in question is genuinely real. The periodic table was confirmed not by a single experiment but by the convergence of electrochemistry, spectroscopy, and atomic mass measurements. General relativity was confirmed by the convergence of perihelion precession, light deflection, and gravitational redshift. When independent observers, using independent methods, find the same geometric structure, the most parsimonious conclusion is that the structure is there.

We are not claiming that the admissibility field has been confirmed with the certainty of general relativity. But the convergence of three independent research programs on the same geometric picture — activation space organized around continuation constraints rather than object categories, structured by level sets, gradient directions, correlation lengths, and flow fields of a single admissibility function — is at minimum a strong suggestion that the right language for activation space is geometric before it is featural, topological before it is categorical, and dynamical before it is static.

The features are not wrong. They are just not fundamental. They are emergent coordinates of a constraint topology that the network has been learning, through billions of prediction tasks, to navigate. The activation manifold is a map of admissible futures. Interpretability, on this view, is not the project of reading off what the map says about the present moment. It is the project of understanding the geometry of the territory the map describes.

References

[1] S. Amari. *Information Geometry and Its Applications.* Springer, 2016.

[2] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds.* Princeton University Press, 2008.

[3] A.-L. Barabási. *Network Science.* Cambridge University Press, 2016.

[4] T. Bricken et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, Anthropic, 2023.

[5] R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

[6] B. Nadler, S. Lafon, R. Coifman, and I. Kevrekidis. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. *Applied and Computational Harmonic Analysis*, 21(1):113–127, 2006.

[7] L. Cunningham et al. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

[8] N. Elhage et al. Toy models of superposition. *Transformer Circuits Thread*, Anthropic, 2022.

[9] Flyxion. *The Admissibility Field*. Zenodo, 2026.

[10] Flyxion. *Frozen Processes: Reachability Ontology and the Geometry of Transformation*. Zenodo, 2026.

[11] Flyxion. *Hidden Manifolds: Projection, Sufficiency, and Reachability in High-Dimensional Systems*. Zenodo, 2026.

[12] Flyxion. *From Preservation to Reachability: Semantic Continuity in an Age of Drift*. Zenodo, 2026.

[13] K. Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010.

[14] D. Ha and J. Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

[15] D. Hassabis and D. Silver. From world models to generalist agents. *Communications of the ACM*, 2024.

[16] J. Holland. *Adaptation in Natural and Artificial Systems*. MIT Press, 1992.

[17] A. Hyvärinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4–5):411–430, 2000.

[18] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley, 2001.

[19] S. Kauffman. *The Origins of Order*. Oxford University Press, 1993.

[20] Y. LeCun. A path towards autonomous machine intelligence. Open Review, 2022.

[21] J. M. Lee. *Introduction to Smooth Manifolds*. Springer, 2013.

[22] S. Liu. Normalization makes activation geometry imaginable. Research Notes, June 9, 2026. https://liusida.com/research-notes/norm_act_geometry/

[23] S. Liu and F. Han. ICA lens: Interpreting language models without training another dictionary. *arXiv preprint arXiv:2606.11722*, 2026.

[24] M. Newman. *Networks*. Oxford University Press, 2018.

[25] C. Olah et al. Zoom in: An introduction to circuits. *Distill*, 2020.

[26] J. Pearl. *Causality*. Cambridge University Press, 2009.

[27] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[28] D. Rumelhart, G. Hinton, and R. Williams. Learning internal representations by error propagation. In *Parallel Distributed Processing*, 1986.

[29] D. Silver et al. Reward is enough. *Artificial Intelligence*, 299, 2021.

[30] S. Stapf, P. Acuaviva Huertos, A. Davtyan, and P. Favaro. World model self-distillation: Training world models to solve general tasks. *arXiv preprint arXiv:2606.12072*, 2026. University of Bern.

[31] D. Stauffer and A. Aharony. *Introduction to Percolation Theory*. Taylor & Francis, 1994.

[32] J. Tenenbaum, V. de Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[33] S. Wolfram. *A New Kind of Science*. Wolfram Media, 2002.

[34] B. Zhang and R. Sennrich. Root mean square layer normalization. *arXiv preprint arXiv:1910.07467*, 2019.