From Classifier to Inverse Engine:

Gesture Recognition as Latent Embodied Semantic Reconstruction

Flyxion

Independent Researcher

May 2026

Abstract

Gesture recognition is conventionally framed as a classification problem over segmented motion events: given an observed hand configuration or trajectory, return the most probable label from a predefined vocabulary. This framing has produced substantial engineering progress, but it systematically misrepresents the structure of the underlying problem. Gesture streams are continuous, coarticulated, user-variable, and semantically underdetermined from any single observational projection. The label is not contained in the gesture; it is reconstructed from the gesture through constraint-compatible inference over learned priors. This review argues that the field is undergoing a deeper structural transformation — from symbolic recognition toward latent embodied semantic reconstruction — and that this transformation is visible simultaneously across wearable inertial systems, vision-language models, swipe keyboard decoding, musical gesture cognition, and multimodal sensor fusion. Drawing on recent advances in ring-type wireless wearables, foundation-model-based gesture classification, and embodied cognition theory, we develop a unified account in which gesture systems are better understood as semantic inverse engines: systems that reconstruct latent embodied intention from partial, temporally extended, multimodal projections over constrained motor manifolds. We interpret the convergence of these subfields through the lens of configurational accessibility and admissibility geometry, and argue that the problem of user-independent generalization — one of the central open challenges in wearable gesture systems — is formally analogous to a sheaf-theoretic gluing problem: the task of constructing coherent global semantic sections from locally variable, signer-specific fiber data. The same structural logic, we argue, underlies swipe keyboard personalization, musical mimetic inference, and trajectory-level tool-use understanding.

Introduction: The Misframing of Gesture Recognition

The dominant paradigm in gesture recognition research begins with a segmentation assumption: that the continuous stream of embodied human action can be decomposed into discrete, bounded units, each of which maps to a semantic label. This assumption is operationally convenient. It reduces a hard continuous inference problem to a more tractable classification problem, permits the use of well-understood supervised learning machinery, and enables clean benchmarking against labeled corpora. It has, accordingly, organized the field for decades, producing substantial advances in isolated gesture classification, static pose estimation, and constrained vocabulary recognition across both vision-based and wearable-sensor architectures [2, 3, 5].

The assumption is also wrong — or rather, it is an approximation that holds well in impoverished experimental conditions and breaks down systematically in the conditions that matter most: continuous signing, natural typing, fluent instrumental performance, tool-mediated skilled labor, and real-time interaction across user populations who were never seen during training. The failure is not incidental. It reflects a structural mismatch between the ontology of discrete classification and the ontology of continuous embodied action.

Human motor behavior is not a sequence of discrete symbolic tokens separated by silent intervals. It is a continuous dynamical process unfolding in a high-dimensional space shaped by anatomical constraints, learned habits, physical affordances, anticipatory planning, and semantic intention. Gestures do not occur in isolation; they deform one another. Every signed word modifies the trajectory of its neighbors. Every keystroke is shaped by the letters that precede and follow it. Every bowing stroke in a musical performance reflects not only the current note but the phrase-level expressive intention that frames it. This property — coarticulation — is not noise to be filtered out. It is itself semantic structure, encoded in the continuous geometry of the action trajectory rather than in the surface identity of discrete tokens [6].

The implications for recognition systems are fundamental. A system that classifies isolated gestures is not merely underperforming on a metric; it is modeling the wrong ontology. The trajectory field is the primary object, and discrete gesture labels are derived abstractions over that field. Recognition systems that proceed in the opposite direction — decomposing continuous input into putatively discrete units before apply-

ing classifiers — are attempting to invert a process that was never discrete to begin with. The errors this introduces are not merely quantitative; they are structural, and they become most visible precisely where performance matters most: at the boundaries between gestures, in the presence of coarticulation, across user populations with different motor habits, and under real-world conditions where the clean segmentation cues present in laboratory data are absent.

This review traces the consequences of this misframing and charts the emerging response to it across several subfields that have historically been treated as independent. The argument proceeds through a sequence of increasingly general theoretical moves. We begin with a survey of the current technical landscape in gesture recognition, organized not as a flat taxonomy but as a progression of approximations toward a more adequate problem formulation. We examine wearable inertial systems — particularly the recent development of ring-type wireless architectures employing gravity-based sensing and foundation-model classification [1] — as the current state of the art in user-independent wearable gesture translation. We then examine swipe-based text input as what we will call the Rosetta Stone of gesture semantics: the earliest mass-deployed system that made the inadequacy of the discrete-classification framing publicly legible, and whose engineering solutions anticipate the theoretical apparatus that more recent systems are only now beginning to formalize. We turn then to musical gesture and embodied cognition, drawing on Arnie Cox's mimetic proxy theory [6] to establish that meaning-bearing gesture interpretation is fundamentally reconstructive and predictive rather than observational, a finding that connects cognitive music theory to the architecture of modern multimodal inference systems in ways that have been largely overlooked. We then extend the analysis to tool use and skilled labor, arguing that the semantics of productive embodied transformation represent the next frontier for the framework we are developing, one that enlarges the significance of gesture recognition from a specialized interface problem into a foundational challenge for embodied AI. Finally, we synthesize these threads into a unified theoretical framework that recasts gesture recognition as a problem of latent action reconstruction over admissibility manifolds, and interprets key open problems in the field — user-independent generalization, sentence-level decoding, multimodal fusion, and personalized motor adaptation — as instances of a single underlying structural challenge.

The unifying thesis is this: gesture recognition systems are converging, across wildly

different engineering substrates and application domains, toward a common architecture. That architecture does not ask "which label matches this input?" It asks: "what latent embodied intention best explains this evolving multimodal trajectory, given the constraints imposed by anatomy, language, habit, physics, and semantic admissibility?" The difference between these two questions is not merely one of formulation. It reflects a fundamental shift in what these systems are trying to do.

A Note on Scope and Method

This paper is a theoretical synthesis grounded in literature review rather than a comprehensive engineering survey. The gesture recognition literature is large enough that exhaustive coverage of all subfields — sign language, affective gesture, pointing, whole-body action recognition, facial gesture, surgical gesture, musical performance analysis — would require multiple volumes. Our coverage is selective, organized by theoretical relevance to the central argument rather than by domain completeness. We prioritize works that illuminate the structural transformation we are describing: the shift from classification to reconstruction, from discrete symbols to continuous trajectories, from user-specific tuning to admissibility-constrained generalization. Technical details of specific architectures are included when they are theoretically informative, not for completeness. Where formal machinery is introduced — particularly in Section 7, which develops the admissibility and sheaf-theoretic interpretations — we attempt to motivate it operationally before naming it, so that readers who do not require the formal apparatus can follow the argument through its engineering instantiations alone.

A Taxonomy of Approaches as a History of Approximations

Vision-Based Systems: Geometry Without Proprioception

The earliest and most extensively developed approach to gesture recognition employs external cameras combined with computer vision algorithms to recover hand pose, trajectory, and identity from image sequences. Vision-based systems have the considerable advantage of being non-contact: they impose no hardware on the user and require no wearable instrumentation. Under controlled laboratory conditions — fixed camera position, uniform background, adequate and consistent illumination — they can

achieve recognition accuracy competitive with the best wearable systems on restricted vocabulary tasks [2, 3, 4].

The limitations of vision-based systems, however, are not merely engineering challenges awaiting better cameras or more powerful models. They reflect a fundamental information-theoretic constraint: the camera captures a two-dimensional projection of a three-dimensional dynamic scene, and this projection discards information that is often essential for gesture disambiguation. Force, proprioception, sub-surface muscular dynamics, and fine-grained finger orientation are all invisible to optical systems. Occlusion — one hand occluding the other, or fingers occluding one another during complex signs — introduces systematic ambiguities that no amount of image processing can fully resolve without additional constraints. Illumination and background variation introduce sensitivity to conditions that are irrelevant to the gesture itself, creating a brittleness that is especially problematic for systems intended to operate in real-world environments rather than controlled laboratory settings [5].

More fundamentally, vision-based systems are external observers of action rather than participants in it. They can recover the geometry of a gesture but not, in general, the proprioceptive and kinesthetic structure that shapes how that geometry is experienced and produced by the signer. This distinction matters for user-independent recognition, because much of the variability between signers is not in the observable geometry of the gesture but in the underlying motor program that produces it. Two signers producing the same sign may generate very similar visual trajectories while exhibiting very different muscular activation patterns, joint torque distributions, and anticipatory preparation movements. A system trained exclusively on visual projections will tend to learn the geometry of the training signers' motor habits rather than the abstract semantic structure of the sign class.

Recent advances in depth sensing (RGB-D cameras, LiDAR), body pose estimation (MediaPipe, OpenPose, and their successors), and video-language models have partially addressed some of these limitations. Depth sensing recovers volumetric geometry inaccessible to RGB cameras alone. Dense pose estimation systems can track dozens of hand landmarks in real time with millimeter-level accuracy under favorable conditions. Video-language models, trained on large-scale paired video and text data, can leverage linguistic priors to disambiguate visually similar gestures in ways that purely geometric classifiers cannot. These developments represent genuine progress, but they do not

dissolve the fundamental projection problem; they add additional projections, each of which recovers different aspects of the latent action while remaining silent about others.

Glove-Based and Wired Wearable Systems

In parallel with vision-based approaches, wearable sensor systems have been developed that measure physical and electrophysiological signals associated with hand gestures directly from the body. The most extensively studied form factor is the instrumented glove, which integrates sensors — strain gauges, flex sensors, inertial measurement units, triboelectric elements, or electromyographic electrodes — into a garment worn over the hand. Glove-based systems offer several advantages over vision-based approaches: they are insensitive to lighting, background, and occlusion; they measure hand state directly rather than inferring it from external projection; and they can capture signals — joint angles, muscular activation, pressure — that are invisible to cameras [7, 8, 9].

The limitations of glove-based systems are, however, substantial, and they bear directly on the central themes of this review. The enclosed form factor reduces breathability and makes prolonged use uncomfortable, particularly during continuous or intensive signing. More importantly, gloves assume fixed sensor placements that do not adapt to individual variations in hand anatomy. Human hands differ significantly in size, finger length, joint position, and skin compliance. A glove designed to place a flex sensor over the proximal interphalangeal joint of a hand of average dimensions will misalign that sensor on hands at either extreme of the size distribution, producing systematic errors in signal acquisition that are not easily corrected by calibration. The result is that glove-based systems tend to perform well for users whose anatomy approximates the design assumption and degrade progressively for users who deviate from it — a form of structural user-specificity that is independent of the classification architecture.

Wired sensor arrays, whether embedded in gloves or mounted on skin surfaces, introduce an additional constraint: physical data transmission lines connecting sensors to a centralized processing unit. As the number of sensing channels increases, the network of transmission wires becomes mechanically constraining, restricting the natural mobility of the hand and interfering with the execution of the complex dynamic ges-

tures that constitute the most challenging and linguistically rich signs. Several studies have documented accuracy degradation in wired systems for high-complexity gestures, attributable at least in part to the mechanical constraint introduced by the wiring [9, 11].

Wireless and Modular Wearable Systems: The Ring Architecture

The most recent generation of wearable gesture systems addresses the wiring constraint through wireless architectures in which each sensor module communicates independently with a host device, eliminating physical data transmission lines between sensors. This shift, while straightforward in principle, has significant practical consequences: it allows each sensor to be positioned independently on the finger, adapting naturally to individual hand anatomy rather than assuming a fixed layout. The modular ring form factor, in particular, allows sensors to be placed on the proximal phalanges of individual fingers, providing a stable anatomical mounting point that accommodates a wide range of hand sizes while maintaining consistent proximity to the finger joints whose motion is most diagnostically significant.

Park et al.'s wirelessly connected, ring-type sign language translator (WRSLT) represents the current state of the art in this architecture [1]. Each ring contains a three-axis accelerometer, Bluetooth Low Energy system-on-chip, power management unit, and compact battery, fabricated on a flexible substrate with serpentine-structured interconnections to accommodate repeated bending during signing. The decentralized architecture — in which each ring establishes an independent Bluetooth peripheral link to the host device, enabling concurrent multilink data acquisition — eliminates not only the wires between fingers but the assumption of centralized processing that conventional wired architectures require.

The sensing strategy is notable. Rather than relying on electromyography, which exhibits large interuser variability due to differences in muscle physiology and electrode-skin impedance, the WRSLT employs gravity-based inertial sensing: the three-axis accelerometer measures both the static projection of gravity under quasi-static conditions (functioning effectively as a tilt sensor) and dynamic acceleration during gesture transitions. This dual capability — simultaneous acquisition of static postural information and dynamic gesture trajectories from a single compact sensor — provides the signal richness needed for recognition while avoiding the user-specificity that biosignal ap-

proaches entail.

On a dataset of 100 American Sign Language words and 100 International Sign Language words, evaluated under strictly unseen-user conditions, the system achieved recognition accuracies of 88.3% and 88.5% respectively [1]. These results are especially significant because they were obtained with only seven sensor rings — selected from an initial full-hand configuration through layer-wise relevance propagation analysis of per-finger contribution to classification accuracy — demonstrating that performance saturation can be reached with a sparse, anatomically optimized sensor configuration rather than full-hand instrumentation.

Biosignal Approaches and Their Structural Limitations

Surface electromyography and related biosignal modalities occupy a distinctive position in the gesture recognition landscape. EMG captures muscular activation patterns that are, in principle, the causal upstream of the hand movements recorded by inertial or optical systems, offering the possibility of detecting gesture intention before it fully manifests as observable motion. In practice, however, EMG-based systems face structural challenges that have proven resistant to engineering solutions. The signal recorded at the skin surface depends not only on the underlying muscular activation but on electrode-skin impedance, subcutaneous fat distribution, electrode placement relative to the muscle belly, perspiration, and individual differences in muscle fiber type composition and recruitment patterns. These factors produce large interuser variability that requires per-user calibration, undermining the user-independence that is a prerequisite for practical deployment [9, 10, 11].

The EMG case is instructive for the theoretical argument of this paper because it illustrates a general principle: signals that are closer to the causal mechanism of action are not automatically more useful for recognition, because they also tend to be more sensitive to the irrelevant individual differences that obscure the semantic content of interest. User-independent recognition requires features that are invariant to biological individuality while remaining sensitive to gestural meaning. Gravity-based inertial sensing achieves this balance by measuring the kinematic consequences of motor programs — the orientations and trajectories that gestures produce — rather than the muscular mechanics that generate them. The kinematic level is sufficiently removed from individual

physiology to be comparatively invariant across users, while sufficiently specific to the gesture to distinguish semantically distinct signs.

Multimodal Fusion as Constraint Multiplication

The various sensing modalities reviewed above are not merely alternative approaches to the same problem; they are complementary projections of the same underlying latent state. Vision captures external geometry but loses force and proprioception. Inertial sensing captures local rotational dynamics but loses scene context and global body configuration. EMG captures muscular intention precursors but is highly user-specific. Audio captures interaction with the environment but is spatially coarse. Each modality is lossy in a different direction; their combination is therefore not merely additive but multiplicative in the constraints it imposes on the latent action state.

This is the correct theoretical framing for multimodal fusion in gesture systems: not feature concatenation, but constraint multiplication over a shared latent manifold. Each modality rules out a different subset of the possible trajectories consistent with the observed data; their intersection converges toward the true latent state far more rapidly than any single modality alone. The practical implication is that modest improvements in any individual sensing modality matter less than intelligent combination of complementary projections. A system with mediocre performance on any single channel can substantially outperform a state-of-the-art single-channel system once the cross-modal constraints are correctly integrated.

This principle already operates implicitly in the WRSLT architecture, where the combination of static gravitational orientation and dynamic acceleration profiles from seven anatomically distributed sensors provides a richer constraint on hand state than any single sensor could. It operates more explicitly in emerging video-language systems, where the combination of visual trajectory information with learned linguistic priors dramatically reduces the ambiguity of continuous signing input. And it operates at its most elaborate in the keyboardless typing and swipe decoding systems that we examine in the following section — systems that achieved practical viability not by improving any single sensing channel but by integrating gesture dynamics with probabilistic lexical knowledge and personalized motor history.

The Swype Inflection Point: Gesture as Probabilistic Trajectory Decoding

In 2003, Cliff Kushler and Randy Marsden filed the patent that would eventually become Swype — a text input system for touchscreens in which the user traces a continuous path through the keys of an on-screen keyboard rather than tapping individual characters. The system reached mass deployment on Android devices around 2010 and was for several years the dominant swipe-based keyboard architecture, used by hundreds of millions of people. Its commercial success was eventually surpassed by Google's keyboard and similar implementations, but its theoretical significance for gesture recognition is undiminished and substantially underappreciated.

The common retrospective account of swipe keyboards frames them as a gesture-to-text mapping problem: the user's finger traces a path, the system recognizes the path as a word. This account is technically accurate and theoretically impoverished. It misses the central innovation, which was not the gesture mapping but the decoding architecture — and specifically the role of personalization in making that decoding architecture work at population scale.

The fundamental problem with swipe text input is that the geometric trajectory is radically underdetermined. Many words produce similar or even nearly identical paths across the keyboard. "There" and "three," for example, pass through nearly the same sequence of key regions, differing only in the precise timing and curvature of the path through the overlapping segments. "Home" and "hone" are more similar still. At the level of raw geometric trajectory, the swipe is frequently insufficient to uniquely identify the intended word without additional information. The path is evidence, not the answer.

What made swipe keyboards practically viable was the integration of that geometric evidence with probabilistic lexical models, language priors, temporal dynamics, biomechanical plausibility constraints, correction history, and — most importantly for the present argument — learned personalized vocabularies. The system's knowledge of what words the individual user commonly employs, how they tend to trace paths for specific words, what corrections they have made in the past, and what linguistic context the current entry occurs in collectively constrain the hypothesis space to the point where the geometric ambiguity becomes tractable. The user's swipe provides a noisy projection of their intended word; the system's accumulated model of that user's mo-

tor habits, vocabulary, and linguistic patterns provides the complementary constraints needed to reconstruct the intended meaning.

This is already a latent action reconstruction system. The observable signal — the touch trajectory on the screen — is a compressed, noisy projection of a latent communicative intention. Reconstruction of that intention requires combining the projection with learned priors over the admissible intention space. The personalized user dictionary is not merely a convenience feature; it is the mechanism by which the system learns the geometry of an individual's motor habits and maps that geometry onto the lexical space in a user-specific way.

The theoretical import of this observation extends well beyond keyboard input. The swipe keyboard system demonstrates, in a domain simple enough to analyze cleanly, the following general principles: first, that observable gesture trajectories are systematically underdetermined with respect to semantic intention; second, that reconstruction of intention from trajectory requires integration of gesture dynamics with learned priors over semantic admissibility; third, that these priors must be personalized to the individual user's motor habits to achieve practical accuracy; and fourth, that the system improves over time through accumulated correction history — a form of online local manifold adaptation that progressively refines the mapping between the individual's motor geometry and the target semantic space.

Each of these principles recurs, in more elaborate form, in every domain this paper examines. The user-independence problem in sign language recognition is the same problem as swipe personalization, viewed from the other direction: instead of learning a user-specific section of the gesture-semantics correspondence, the system must learn a representation general enough that individual users' sections are all locally consistent with a common global section. The sentence-level sliding-window decoding in the WRSLT system is an elementary version of the language-prior integration that made swipe decoding tractable. The mimetic inference that Cox identifies in musical gesture perception is a more elaborate instance of the same principle: reconstruction of latent embodied intention from partial, noisy, projection-mediated traces, using learned models of admissible action trajectories.

Coarticulation and the Failure of Isolated Classification

The inadequacy of isolated gesture classification becomes most visible when systems are exposed to continuous streams of embodied action rather than carefully segmented laboratory examples. In such settings, the assumption that gestures exist as discrete, bounded entities separated by neutral intervals rapidly collapses. Human motor behavior is profoundly coarticulated: each action is shaped by the actions that precede and follow it. The trajectory executed at time t is not solely a function of the semantic target currently being expressed, but of anticipatory preparation for future states, residual momentum from prior states, biomechanical optimization, energetic minimization, and higher-order phrase-level structure.

Coarticulation is most familiar in speech production, where phonemes are not realized as invariant acoustic objects but as context-sensitive deformations shaped by neighboring articulatory targets. The pronunciation of a consonant changes depending on the vowels surrounding it; the acoustic realization of a syllable depends on its position within the phrase; prosody propagates continuously across nominal token boundaries. Crucially, these effects are not imperfections contaminating an otherwise discrete symbolic process. They are constitutive of fluent speech itself. Human listeners do not decode speech by isolating acoustically invariant phoneme tokens because such tokens do not exist in natural speech. Rather, listeners reconstruct the latent articulatory and semantic trajectory responsible for the continuously evolving acoustic field.

The same structural logic governs handwriting, typing, sign language, and musical performance. In handwriting, the geometry of a letter depends on the letters surrounding it, because the hand optimizes continuously over the trajectory rather than independently rendering isolated glyphs. In typing, finger trajectories reflect anticipatory preparation for future keys, producing timing patterns and motor shortcuts that are user-specific yet highly systematic. Swipe keyboards make this visible at the geometric level: the path connecting letters is shaped by the entire intended word rather than by local transitions alone. The trajectory is globally optimized under motor and lexical constraints.

Sign language exhibits coarticulation even more dramatically because the body itself functions as the articulatory substrate. Signed words blend continuously into one another through transitional handshapes, anticipatory wrist rotations, residual arm mo-

mentum, and phrase-level spatial structuring. The hand does not return to a neutral origin state after each sign; it evolves continuously through a constrained trajectory field whose local structure depends on future semantic targets. The practical consequence is that the temporal boundaries between signs are often indeterminate even for expert human annotators. This ambiguity is not a failure of annotation. It reflects the fact that the segmentation itself is not ontologically primary.

The WRSLT system discussed earlier provides a revealing illustration of this problem [1]. The system's sentence-level translation architecture avoids direct training on complete sentence trajectories, instead using a sliding-window framework in which words are detected only after repeated appearance across overlapping temporal windows. This mechanism is significant because it implicitly acknowledges that semantic identity is temporally distributed rather than instantaneously observable. A gesture candidate is not accepted because a single frame or local segment uniquely identifies it. It becomes stabilized only after persistence across multiple partially overlapping projections.

This persistence criterion can be interpreted as a primitive form of trajectory coherence filtering. Candidate interpretations that fail to maintain consistency across adjacent temporal windows are treated as transitional artifacts rather than semantically stable states. The system therefore reconstructs meaning not from isolated local observations but from the persistence structure of the evolving trajectory itself. Although implemented operationally as a sliding-window heuristic, the underlying principle is substantially deeper: semantic admissibility is established through temporal coherence over overlapping projections. The trajectory is evidence, not the meaning. And evidence must accumulate before a stable semantic commitment can be made.

The broader implication is that discrete classification systems fail not because they are insufficiently powerful classifiers, but because they presuppose the existence of stable local tokens where the underlying motor process contains only dynamically evolving constraint structures. The apparent success of isolated classification benchmarks is partly an artifact of dataset construction. Most benchmark corpora artificially segment gestures before training and evaluation, thereby externalizing the segmentation problem and presenting the classifier with already-discretized objects. Real-world gesture streams do not arrive pre-segmented. The segmentation itself is an inference problem, and it cannot be cleanly separated from the recognition problem it is supposed to pre-

cede.

There is a further dimension of coarticulation that bears on the ambiguity structure of gesture streams: what we may call dynamical ambiguity, which is distinct from the geometric and semantic ambiguities discussed in earlier sections. Geometric ambiguity arises when multiple gestures produce similar spatial configurations. Semantic ambiguity arises when the same gesture maps to multiple plausible interpretations in the lexical space. Dynamical ambiguity arises when identical local geometries belong to radically different trajectory-level regimes depending on temporal context, anticipatory structure, or phrase-level intent. A hand configuration that appears identical in a single frame may be part of a transition into gesture A in one context and a transition out of gesture B in another. The geometry does not determine the meaning; the geometry's position within the larger dynamical trajectory is what matters. Classifiers trained on local segments cannot, in principle, resolve this ambiguity.

This observation also helps explain a pattern that has been noted across multiple areas of gesture recognition: systems that achieve high accuracy on isolated benchmark tasks frequently degrade substantially in continuous real-world conditions. The degradation reflects not overfitting to geometric detail but overfitting to the implicit discretization structure of the training data. The classifier has learned the geometry of already-segmented gestures; when presented with continuous streams where segmentation must be inferred, it encounters objects it was not trained to handle.

The field's response to this failure has been the gradual introduction of sequence modeling architectures that represent temporal context rather than local states. Recurrent neural networks, long short-term memory architectures, and transformer-based sequence models have all been applied to gesture recognition with the explicit motivation of capturing the temporal dependencies that local classifiers miss. This progression is significant. It represents a partial empirical discovery of the coarticulation principle: that semantic content is distributed over time rather than localized in instantaneous states. The architectural evolution from frame-level classifiers to sequence models to sliding-window coherence frameworks tracks the progressive incorporation of trajectory-level structure into recognition systems.

Coarticulation therefore represents more than a technical nuisance. It reveals a deeper ontological fact about embodied communication: semantic structure is distributed across continuous trajectories rather than localized in discrete motor events. Any ade-

quate theory of gesture recognition must therefore operate primarily over evolving trajectory manifolds rather than over isolated symbolic classes. The classifier paradigm survives only as a local approximation imposed on a fundamentally continuous process, and its failures become most visible precisely where the approximation is worst: at boundaries, under real-world conditions, and across user populations whose motor habits differ from those represented in the training corpus.

Music, Mimetic Cognition, and the Reconstruction of Embodied Intention

The problem of reconstructing latent embodied intention from partial trajectories is not unique to gesture interfaces or sign language recognition. It appears with particular clarity in music cognition, where listeners routinely infer bodily gesture, physical effort, expressive intention, and emotional structure from acoustic signals alone. Musical perception is therefore not merely auditory. It is deeply embodied, involving covert simulation of the bodily actions capable of producing the observed sound structures.

Arnie Cox's mimetic hypothesis provides one of the most developed accounts of this phenomenon [6]. According to Cox, musical understanding depends fundamentally on mimetic participation: listeners internally simulate or imaginatively approximate the bodily actions associated with musical production. Hearing a violin phrase, for example, may recruit covert motor representations related to bow pressure, arm trajectory, fingering tension, and dynamic articulation even in listeners who are not violinists themselves. The listener does not merely decode an abstract acoustic pattern. They reconstruct, implicitly and partially, the embodied action structure capable of generating that pattern.

The importance of this theory for gesture recognition is difficult to overstate. It implies that human interpretation of dynamic sensory signals already operates as a form of latent inverse inference. Meaning is not extracted directly from observable surface structure. It is reconstructed through constraint-compatible simulation of plausible underlying generative processes. Human listeners infer intention by implicitly modeling the space of admissible bodily trajectories capable of producing the observed sensory trace.

This logic parallels the architecture increasingly visible in modern gesture systems. A sign-language recognizer does not simply map hand geometry onto lexical labels.

It reconstructs the most plausible latent communicative trajectory consistent with the observed multimodal evidence. A swipe keyboard does not merely classify geometric paths; it infers intended lexical structure from noisy motor traces constrained by language priors and personalized usage history. A multimodal wearable system does not observe meaning directly; it progressively narrows the admissible action manifold until a semantically stable reconstruction becomes possible.

Music provides an especially revealing case because expressive musical gestures often remain intelligible even when the underlying sensory channel is highly impoverished. A listener can frequently infer tension, release, hesitation, effort, aggression, delicacy, or anticipation from sparse acoustic cues alone. This suggests that the interpretive system is not operating through direct feature matching but through reconstruction over deeply learned embodied priors. The projection is evidence, not the intention. And the prior model of admissible embodied gesture is what makes the reconstruction tractable despite the projection's incompleteness.

The relationship between musical phrasing and coarticulation is particularly significant. In fluent musical performance, individual notes are rarely produced as isolated acoustic events. Bow trajectories, breathing patterns, fingering transitions, vibrato envelopes, and timing microvariations propagate continuously across phrase boundaries. The identity of any local gesture depends strongly on its position within the larger expressive trajectory. A note performed identically at the acoustic level may convey radically different meaning depending on phrase-level context, anticipatory tension, or dynamic shaping. The same dynamical ambiguity that complicates sign language decoding appears here at a finer temporal scale and with expressive rather than purely semantic consequences.

Musical performance also reveals the importance of physical admissibility constraints in shaping embodied trajectories. Instrumental technique occupies a constrained manifold determined by anatomy, instrument geometry, energetic efficiency, learned motor schemas, and stylistic convention. A violinist's bowing motions, a pianist's fingering transitions, or a singer's breath control patterns are not arbitrary movements through state space. They are highly constrained trajectories optimized under interacting biomechanical and expressive pressures. The admissibility landscape of a trained performer is not the same as that of a novice, and is not the same across instruments or stylistic traditions. But it is always a structured landscape rather than an unconstrained space,

and this structure is precisely what makes gesture interpretation tractable.

This observation connects to a broader theoretical point about the relationship between personalization and admissibility. In swipe keyboard systems, personalized dictionaries encode the individual user's motor habits and lexical preferences. In sign language recognition, user-independent generalization requires learning an admissibility structure that accommodates individual variation while remaining semantically coherent. In music cognition, mimetic reconstruction operates over prior models shaped by the listener's own motor experience and cultural familiarity with musical idioms. In all three cases, the quality of interpretation depends on the quality of the prior model of admissible action trajectories. Systems that lack such models are forced to treat the full state space as uniformly plausible, which drowns the useful signal in combinatorial noise.

Cox's mimetic framework also helps explain why multimodal systems increasingly outperform purely geometric classifiers. Human interpretation of embodied action rarely relies on a single sensory projection. Visual observation, proprioceptive simulation, acoustic coupling, contextual expectation, and prior experience collectively constrain the inferred latent action. Modern AI systems appear to be converging toward a similar architecture. Video-language models, multimodal wearables, and trajectory-based gesture systems increasingly operate by integrating complementary projections over shared latent action manifolds rather than by performing isolated classification on single-channel inputs.

The mimetic hypothesis also points toward a further dimension that gesture recognition systems will need to address as they mature: the distinction between semantic gesture and expressive gesture. Much of the gesture recognition literature focuses on semantically discrete communicative acts — signs with defined lexical meaning, typed words with defined orthographic identity, commands with defined functional referents. Musical gesture, however, is often irreducibly expressive: it conveys affect, effort, dynamic shaping, and structural articulation that are not reducible to a discrete semantic label. The gesture is the meaning, not a vehicle for a separate propositional content. This suggests that mature embodied inference systems will need to operate in a richer semantic space than current lexical classification architectures provide — a space that encompasses continuous expressive dimensions alongside discrete communicative functions.

From this perspective, the recent evolution of gesture recognition systems begins to

resemble a technological rediscovery of principles long operative in embodied human cognition. The trajectory itself is never fully observed. What is observed are partial projections constrained by prior models of admissible embodied action. Interpretation emerges through reconstruction, stabilization, and prediction across these incomplete projections. Gesture systems are therefore evolving not toward better symbolic classifiers, but toward increasingly sophisticated engines for inverse embodied semantic inference.

Tool Use, Skilled Labor, and the Semantics of Embodied Transformation

The theoretical framework developed in the preceding sections — gesture as continuous trajectory in a constrained admissibility landscape, interpretation as latent inverse inference over partial projections — becomes even more significant when extended beyond communicative gesture into productive action. Hammering, suturing, soldering, painting, carpentry, weaving, cooking, and instrument repair are not communicative acts in the ordinary sense. They do not express propositions or lexical items. They enact physical transformations in the world. Yet they share with sign language and musical performance the same fundamental structure: continuous trajectories through constrained motor manifolds, shaped by task physics, material affordances, anatomical biomechanics, learned habits, and semantic intention.

This extension is important because it reveals that gesture recognition is not merely a specialized interface problem. It is the surface expression of a broader scientific challenge: understanding how embodied agents navigate structured action landscapes to produce semantically meaningful physical effects. That challenge is central to robotics, surgical training, industrial skill transfer, rehabilitation medicine, augmented reality guidance, and the development of embodied AI systems capable of learning from human demonstration. All of these domains require systems that can infer not merely what motion is occurring but what productive intention that motion is serving, at what stage of execution, and with what degree of proficiency.

The keyboard provides a useful transitional example. A keyboard is an external discretization scaffold: it collapses a continuous motor trajectory into a sequence of explicit symbolic events — key down, key up, ordered character sequence — by physically discretizing the space of hand configurations. The keyboard does not recognize the user's

intention; it provides a physical interface that makes intention legible by constraining the space of possible outputs. Keyboardless typing, by contrast, removes this scaffold and forces reconstruction from continuous projection alone. The finger trajectories are continuously optimized under linguistic intention and motor habit but never forced through discrete affordances. A system capable of keyboardless typing must therefore solve the same inverse inference problem that sign language recognition and swipe decoding require: recovering the latent semantic structure from the unconstrained motor trajectory.

This framing reveals that the perceived difficulty of keyboardless typing relative to standard keyboard use is not primarily a matter of motor skill. The human motor system can execute precise, repeatable finger trajectories in the absence of physical key affordances; typists who have learned to type on glass surfaces often maintain near-standard speeds. The difficulty is epistemic: without the physical discretization that keyboards provide, the system observing the action must perform the segmentation and recognition internally. The keyboard outsourced this inference to mechanical constraints; keyboardless systems must perform it computationally.

The same logic applies across the full range of skilled tool use. A skilled craftsperson hammering a nail does not execute a stereotyped ballistic movement. They continuously adapt force, velocity, angle, and trajectory based on real-time feedback from the nail, the material, the tool, and proprioceptive information about hand position and grip pressure. The action is semantically structured — it aims at a transformation state that the agent is trying to bring about — but that semantic structure is distributed across the entire trajectory and does not reside in any local segment. An observer attempting to classify the craftsperson's action from a single frame or a short local window will consistently fail for the same reasons that isolated-gesture sign language classifiers fail: the semantic information is in the trajectory structure, not in instantaneous states.

Industrial skill tracking, surgical gesture recognition, and rehabilitation monitoring have begun to address these problems [14], but the field remains primarily organized around action segmentation and categorical classification rather than trajectory-level reconstruction. The dominant architecture — segment the action stream into approximately discrete events, classify each event against a predefined vocabulary — imports the same ontological assumption that hobbles sign language classifiers. It works adequately for coarse-grained action categories in controlled settings and degrades sub-

stantially in fine-grained, continuous, naturalistic conditions.

The transition toward trajectory-level reconstruction in tool-use domains will likely require the same theoretical moves that are already underway in sign language and musical gesture systems: integration of multimodal constraint sources, learning of domain-specific admissibility manifolds, continuous sequence modeling that captures coarticulation structure, and personalization mechanisms that adapt to the individual practitioner's motor habits. The challenge is substantially harder in tool-use domains because the admissibility landscape is defined not only by anatomy and language but by the physical properties of materials, the geometry of tools, and the state of the work in progress. These constraints are highly context-specific and often only partially observable.

What this analysis suggests is that tool-use inference systems must eventually incorporate models of the physical world as well as models of the agent's motor habits. The question "what is this person doing?" requires not merely a model of how people move but a model of what physical transformations their movements are capable of producing, given the tools and materials present. This is a substantially more demanding inverse inference problem than gesture lexicon recognition, but it is not categorically different in structure. It involves recovering latent embodied intention from partial, multimodal, temporally extended observations, constrained by prior models of admissible action in a given physical and semantic context. The semantic inverse engine required for tool-use understanding is the same architecture as the one converging toward in sign language, swipe decoding, and musical gesture interpretation — only the admissibility landscape is richer and the semantic targets are physical states rather than lexical items.

This generalization has a further implication worth noting: the distinction between tool use, sign language, and musical performance may be less fundamental than the commonality of their underlying inference structure. All three are forms of skilled embodied transformation: transformation of physical materials in the case of tools, transformation of a communication channel in the case of sign language, transformation of acoustic fields in the case of music. All three require continuous navigation of constrained motor manifolds under semantic intention. All three produce observable trajectories that are partial, noisy projections of latent action structures that can only be recovered through constraint-compatible reconstruction. The gesture recognition prob-

lem, broadly construed, is simply the inverse problem of skilled embodied agency.

Toward a Unified Framework: Latent Action Reconstruction over Admissibility Manifolds

The preceding sections have accumulated a set of recurring structural observations across domains that are conventionally treated as independent: wearable gesture recognition, swipe keyboard decoding, musical mimetic cognition, and tool-use inference. In each domain, the observable signal is a partial, noisy projection of a latent embodied state. In each domain, the projection is underdetermined: multiple latent states are consistent with any local observation. In each domain, disambiguation requires integration of the partial observation with prior knowledge of what trajectories are admissible — consistent with anatomy, physics, language, convention, and accumulated individual habit. In each domain, interpretation is a temporal process that stabilizes over time rather than resolving instantaneously. And in each domain, the field is converging, through engineering necessity, toward architectures that perform this kind of constraint-accumulating, temporally extended, multimodal reconstruction.

This section develops a unified account of these observations and argues that they are not coincidental convergences but expressions of the same underlying mathematical structure.

The Projection Problem and the Admissibility Landscape

Let us make precise the structural claim that every sensing modality provides a partial projection of a latent action state. A gesture, in the most general sense, is a trajectory through a high-dimensional state space parameterized by joint angles, velocities, contact forces, muscular activation levels, proprioceptive fields, and the agent's evolving semantic intention. Call this space $\mathcal{M}$, the motor manifold. An observation from any single sensing modality — a camera image, an accelerometer reading, an EMG envelope, an acoustic recording — is a function $\pi_i : \mathcal{M} \to \mathcal{O}_i$ mapping the latent motor state to an observation in modality i's observation space $\mathcal{O}_i$. Because each π_i discards information — vision discards force, inertial sensing discards global context, EMG discards external geometry — the preimage $\pi_i^{-1}(o_i)$ of any single observation o_i is a submanifold of $\mathcal{M}$ rather than a point. The observation constrains the latent state but does not determine

it.

Disambiguation requires additional constraints. These constraints come from two sources: the structure of the observation itself as it evolves over time, and prior knowledge of which regions of $\mathcal{M}$ are admissible. We formalize this as follows.

Definition (Admissibility operator). Let $C_t : \mathcal{M} \to \{0, 1\}$ be a time-dependent admissibility operator encoding the anatomical, physical, linguistic, energetic, and habitual constraints governing the agent and the task at time t. The *admissibility landscape* at time t is

$$\mathcal{A}_t = \{x \in \mathcal{M} \mid C_t(x) = 1\}.$$

Not all trajectories through $\mathcal{M}$ are admissible. C_t encodes the structure of the accessible region and evolves as the agent's context, state, and intentions evolve. This is the primary object of the framework: later appendices formalize projection, coarticulation, gluing, and temporal stabilization as operations on or properties of $\mathcal{A}_t$ and the trajectory families it admits.

Definition (Trajectory measure). Admissible trajectories are not merely geometrically possible; they are weighted by plausibility under energetic, linguistic, and habitual constraints. Define a trajectory functional

$$P(\gamma) \propto e^{-\mathcal{S}[\gamma]},$$

where $\mathcal{S}[\gamma]$ is an action-like constraint functional penalizing trajectories that are unlikely under prior models of motor habit, linguistic expectation, and physical efficiency. The trajectory measure P assigns higher weight to trajectories near attractor basins of the admissibility landscape and lower weight to trajectories near its boundaries. Recognition is then not the identification of the geometrically unique trajectory consistent with evidence, but the inference of the trajectory of maximum posterior weight consistent with accumulated projections and admissibility constraints.

Recognition, in this framework, is the problem of identifying the trajectory $\gamma \in \mathcal{A}_t$ of highest $P(\gamma)$ consistent with the accumulated multimodal evidence. Each new observation $o_i(t)$ from modality i at time t constrains the set of admissible trajectories consistent

with the evidence to date. The intersection of these constraints across modalities and time progressively concentrates the trajectory measure on a small region of $\mathcal{A}_t$ until the reconstruction stabilizes at a semantically determinate state. The gesture label is not read off from the trajectory; it is the name of the admissible region of $\mathcal{A}_t$ toward which the posterior measure converges.

This formulation makes explicit several features that were implicit in the engineering systems examined earlier. Multimodal fusion is not feature concatenation; it is constraint multiplication. Each modality contributes a projection constraint that, in combination with others, concentrates the trajectory measure on a progressively smaller region of $\mathcal{A}_t$. The WRSLT's combination of static gravitational orientation and dynamic acceleration across seven anatomically distributed sensors is an instance of this: seven correlated partial projections that jointly constrain the accessible motor state far more tightly than any single accelerometer could. The sliding-window persistence criterion is an instance of temporal constraint accumulation: the recognition system requires that the trajectory remain in a consistent region of $\mathcal{A}_t$ across overlapping temporal windows before committing to a semantic interpretation. Swype personalization is an instance of individual admissibility refinement: the personalized dictionary encodes a user-specific estimate of which regions of $\mathcal{A}_t$ are accessible to a particular agent, refining the global prior with individual motor history.

Hierarchical Motor Oscillators and the Generative Structure of Gesture

The admissibility framework specifies which trajectories through $\mathcal{M}$ are semantically admissible, but says comparatively little about how those trajectories are generated in the first place. This gap matters because the generative structure of embodied motion is not arbitrary. It imposes further constraints on the shape and topology of $\mathcal{A}_t$ that the inference layer must respect to be biologically realistic. The missing middle layer between semantic intention and observable motion is supplied by hierarchical coupled oscillatory dynamics, or central pattern generator (CPG)-like architectures, which modern motor neuroscience increasingly recognizes as the organizational substrate of fluent embodied action across locomotion, speech, handwriting, sign language, musical performance, and tool use.

A gesture is not a discrete symbolic emission but a coordinated excitation pattern

propagating through nested motor oscillatory assemblies. The essential insight is that higher-level semantic intention does not specify trajectories point-by-point. Rather, it biases or modulates the regime of lower-level oscillatory assemblies whose emergent coordination produces the observable trajectory stream. The resulting motion is intrinsically dynamical: generated by phase relationships, entrainment, inhibition, and local correction dynamics rather than by explicit trajectory commands.

This generative structure immediately explains several phenomena the paper has discussed on phenomenological grounds. Coarticulation is not an imperfection superimposed on discrete token emission; it is intrinsic to oscillator entrainment dynamics. The motor oscillators serving a future gesture begin to entrain before the current gesture completes, because coupling propagates continuously through the oscillator network rather than waiting for symbolic boundaries. Residual phase structure from preceding gestures persists after their nominal completion for the same reason. Gesture boundaries therefore become soft synchronization transitions — phase-locking events in a continuously evolving coupled system — rather than discrete symbolic separations. The ambiguity of segmentation is not a failure of annotation; it reflects the fact that phase transitions in nonlinear coupled oscillators are inherently continuous.

User variation also receives a natural mechanistic account. Different individuals instantiate the same semantic gesture attractor through different oscillator coupling geometries, different intrinsic frequencies, different entrainment time constants, and different local correction thresholds. The semantic target — the high-level phase-locking configuration corresponding to a particular sign, phoneme, or musical phrase — may be shared across users while the particular trajectory through $\mathcal{M}$ that achieves it varies systematically with individual motor geometry. This is exactly the structure of local sections in a sheaf: locally consistent realizations of a shared global semantic object, varying in details determined by individual oscillator architecture.

The connection to the hippocampal subspace rotation results discussed in Section 9 is also clarified by this framework. The rotating communication subspaces can be reinterpreted as dynamically reconfigurable coupling geometries between oscillatory subsystems: the same neural population participates in different functional coordination patterns depending on which phase relationships are currently locked. Context-dependent computation becomes context-dependent oscillatory entrainment, achieved through geometric reorientation of the communication subspace rather than anatomical rewiring.

For musical gesture, the CPG framework is especially apt. Meter, subdivision, bowing cycles, breathing patterns, vibrato, and phrase-level tension operate through exactly this kind of nested synchronization hierarchy. The expressive dimensions of musical gesture — timing microvariations, dynamic shaping, articulation — are perturbations of the oscillatory equilibrium imposed by technical and expressive intention, and the listener's mimetic reconstruction of them (as described by Cox's mimetic hypothesis) can be understood as implicit inference over the generative oscillatory structure underlying the observable acoustic trajectory. The mathematical framework that describes CPG-driven gesture generation and the active-medium inference framework that describes gesture reconstruction are thus not merely analogous: they are dual descriptions of the same embodied dynamical system.

The admissibility landscape $\mathcal{A}_t$ can now be given its most precise interpretation: it is the set of trajectories in $\mathcal{M}$ arising from dynamically stable phase-locked configurations of the hierarchical oscillator network. Admissible trajectories are not simply geometrically possible paths but dynamically realizable ones — trajectories that can be sustained by the coordinated attractor dynamics of the motor oscillatory system. Stability under perturbation corresponds to positive accessibility curvature (Appendix F): gesture classes whose semantic attractor basin is deep and broad in the oscillator phase space will be robust to individual variation and environmental noise. Classes whose attractor basin is shallow or narrow will be fragile, requiring precise coordination that is easily disrupted.

The formal development of hierarchical CPG-like systems and their relationship to the admissibility and coarticulation frameworks is continued in Appendix B, where the coupling equations and trajectory emergence are specified and their implications for the Projection Insufficiency theorem and ontological stability of local classification are drawn out.

The Unseen-User Problem as a Sheaf-Theoretic Gluing Condition

The problem of user-independent gesture recognition deserves special attention because it is both practically central and theoretically revealing. Most gesture systems that achieve high accuracy under seen-user conditions degrade substantially when evaluated on users who were not represented in the training data. This degradation is con-

ventionally attributed to domain shift, overfitting, or insufficient representation robustness, and addressed through data augmentation, domain adaptation techniques, and larger training sets. These are useful engineering responses, but they do not fully capture the structural nature of the problem.

The structural nature becomes clear when we consider what user-specific variation actually is. Each signer, typist, or performer occupies a distinctive region of the motor manifold. The gesture class "want" in American Sign Language is not a single fixed trajectory but an equivalence class of trajectories, one for each signer, that share the relevant semantic and linguistic features while differing in the details determined by individual anatomy, habit, and motor history. Each signer's realization of the class is a *local section* of a gesture-semantics correspondence: a locally consistent mapping between motor trajectories in that individual's motor geometry and semantic labels. The sections are locally consistent — each signer's signing is recognizable to other signers and to trained classifiers — but they are not globally identical. The motor geometry varies across signers in ways that a purely geometric classifier, trained on a limited number of individuals, may not generalize across.

User-independent recognition therefore requires constructing a *coherent global section* from a collection of locally consistent but geometrically distinct signer-specific sections. This is precisely the structure of a sheaf-theoretic gluing problem. A sheaf over a base space assigns local data to open sets and provides gluing conditions that determine when locally consistent local sections can be assembled into a globally consistent global section. In the present context, the base space is the space of gesture classes, the local data are signer-specific motor trajectories, and the gluing condition is the requirement that local sections be compatible on overlaps — that two signers' realizations of the same gesture class are sufficiently similar in the shared latent space that a common section can be defined.

When this gluing condition holds, a user-independent system can be constructed: the global section defines a recognition function that maps arbitrary new signers' trajectories to the correct semantic class without per-user calibration. When the gluing condition fails — that is, when local sections are locally consistent but globally incompatible — user-independent recognition is structurally impossible without additional constraints, and the failure corresponds to a non-trivial obstruction class in the Čech cohomology of the gesture manifold.

The practical importance of this analysis is that it identifies the correct target for user-independent system design. The goal is not merely to train on more users or to apply standard augmentation techniques. The goal is to learn a semantic embedding space in which the gluing condition is approximately satisfied: a space where the local sections contributed by different signers are mutually consistent, so that a coherent global section exists. The WRSLT's semantic prototype alignment — projecting sensor embeddings into a shared latent space organized by text embeddings of word meanings — is a direct engineering attempt to construct such a space. By aligning gesture embeddings with semantic word embeddings across users, the system attempts to learn a representation in which inter-user geometric variation is factored out of the semantic structure, leaving a shared embedding geometry where local sections from different signers can be coherently assembled.

The success of this approach under unseen-user conditions — 88.3% accuracy on 100 ASL words and 88.5% on 100 ISL words — provides empirical evidence that the gluing condition is approximately satisfiable in the learned embedding space, at least for the gesture classes and user population studied. The remaining error reflects the degree to which the gluing condition fails: cases where individual signer variation is sufficiently large that local sections cannot be coherently assembled in the learned representation.

Temporal Stabilization as Constraint Closure and Irreversible Commitment

The temporal dimension of gesture recognition adds a further layer to the admissibility framework. Recognition is not a static classification at a single moment but a dynamic process of accumulating evidence over time. As the trajectory evolves, successive observations progressively constrain the accessible volume of the admissibility landscape until the reconstruction stabilizes at a determinate semantic state. This process has the structure of a constraint closure: starting from a large set of admissible hypotheses, each new observation eliminates hypotheses inconsistent with the accumulated evidence until a closed, stable interpretation is reached.

This process is irreversible in a significant sense. Once the accessible volume has been reduced to a single semantic region and the system has committed to an interpretation, retreating from that commitment to entertain previously eliminated hypotheses requires structural revision rather than merely updating an estimate. The commitment

event — the moment at which the trajectory stabilizes into a determinate semantic state — is not a probabilistic update in the ordinary Bayesian sense. It is a qualitative change in the epistemic status of the trajectory: from candidate to confirmed, from admissible to actualized.

In the WRSLT sentence-level framework, this commitment event corresponds to the moment at which a word candidate has appeared persistently across the required number of consecutive sliding windows. Before that threshold, the word is a candidate: the trajectory is in the admissible region for that class, but the evidence is insufficient for commitment. After the threshold, the word is confirmed: the trajectory has maintained sufficient coherence across overlapping temporal projections to justify irreversible semantic commitment. The boundary is sharp in the system's implementation, though the underlying continuous trajectory is, of course, smooth.

This event-like structure of semantic commitment maps directly onto the logic of event calculi in formal semantics, and more specifically onto the Kinetic-Event Synthesis framework in which irreversible observational commitments of the form $\Omega_t \to H_{t+1}$ represent the passage from potential to actualized semantic content. The gesture trajectory evolves continuously through the admissible region Ω_t; the commitment event collapses this to a determinate historical record H_{t+1} that forms the basis for subsequent interpretation. Each confirmed word in the sentence detection framework is such an event: an irreversible stabilization of a portion of the continuous trajectory into a fixed semantic record that constrains the interpretation of what follows.

Personalized Motor Dictionaries as Local Manifold Adaptation

The fourth major structural theme that the unified framework must accommodate is personalization: the systematic variation in gesture-semantics correspondence across individuals, and the improvement in recognition performance that follows from learning individual-specific parameters. We have already discussed this in the context of swipe keyboards and the sheaf-theoretic analysis of user-independent generalization. Here we develop the point more explicitly in terms of the admissibility landscape.

An individual's motor history defines a personal admissibility landscape: a structured estimate of which trajectories are accessible, efficient, and semantically reliable for that individual. This landscape is shaped by the individual's anatomy, learned

motor habits, linguistic experience, and cultural familiarity with the gesture vocabulary. It differs systematically from the population-average admissibility landscape that a user-independent system must approximate. The discrepancy between individual and population-average landscapes is the source of the performance gap between seen-user and unseen-user conditions observed across virtually all gesture recognition systems.

Personalized adaptation corresponds to learning an individual's local admissibility landscape and using it to refine the recognition function. In swipe keyboards, this adaptation is implemented through correction history and personalized vocabulary. In few-shot learning systems, it corresponds to fine-tuning on a small number of individual-specific examples. In the WRSLT architecture, the fixed sensor placement on the proximal phalanges provides a degree of anatomical normalization that reduces individual variation without requiring explicit per-user calibration, an engineering approximation to local admissibility alignment.

The theoretical ideal is a system that maintains both a global model of the population-level admissibility landscape — enabling generalization to new users — and a mechanism for rapid local adaptation to individual-specific deviations from the population model. This dual architecture is already implicit in the most successful gesture recognition systems: a strong prior over population-level gesture structure, updated by individual-specific evidence as it accumulates. The convergence of this architecture across sign language, swipe decoding, musical gesture, and tool-use inference suggests that it reflects a genuine structural requirement of the inverse reconstruction problem rather than a contingent engineering choice.

The Convergence: Gesture Systems as Semantic Inverse Engines

The preceding analysis has traced a convergent movement across multiple subfields of gesture recognition and embodied action understanding. Vision-based systems, wearable inertial systems, biosignal approaches, swipe keyboards, musical gesture analysis, and tool-use inference have all, through independent engineering necessity, been pushed toward architectures that share a common functional structure: they do not classify gestures against a fixed vocabulary of predefined labels. They reconstruct latent embodied intentions from partial, temporally extended, multimodal projections

over constrained action manifolds.

This convergence is not accidental. It reflects the structure of the problem itself. Embodied action is continuous, coarticulated, user-variable, and semantically underdetermined from any single observational projection. Systems that attempt to circumvent these properties through clever engineering — larger cameras, more sensitive EMG, faster processors, bigger training sets — can delay but not eliminate the failure modes that these properties produce. Only systems that are structurally designed to perform inverse embodied inference can address the problem at its root.

The functional architecture of such systems has several characteristic features. First, they treat observable signals as evidence rather than as direct semantic content. The trajectory is evidence, not the meaning. The projection is evidence, not the intention. This epistemic reorientation transforms the recognition problem from a classification task into an inference task, and opens the door to the full apparatus of probabilistic inference, constraint accumulation, and Bayesian updating.

Second, they integrate multiple partial projections as complementary constraints over a shared latent manifold rather than concatenating features or voting across channels. This is the constraint-multiplication principle: each additional modality narrows the admissible volume rather than adding information to an independent count. The practical consequence is that modest improvement in integration architecture often yields larger performance gains than substantial improvement in any single sensing modality.

Third, they model the structure of the admissibility landscape rather than merely learning the geometry of training-data trajectories. The admissibility landscape encodes what trajectories are possible and meaningful, not merely what trajectories have been observed. Systems that learn this structure generalize to new users, new contexts, and new vocabularies in ways that systems trained only on observed trajectory geometry cannot.

Fourth, they perform temporally extended reconstruction that accumulates evidence over time rather than committing to interpretations at each local instant. Semantic stabilization is a process, not an event. The commitment threshold may be implemented as a hard decision rule, as in the WRSLT sliding-window framework, but the underlying process is continuous accumulation of constraint-compatible evidence.

Fifth, they maintain mechanisms for individual adaptation that allow the global

model to be locally refined for specific users. The tension between generalization and personalization is not a dilemma to be resolved by choosing one at the expense of the other; it is a structural feature of the problem that requires simultaneous modeling at both population and individual levels.

These five features collectively define the semantic inverse engine architecture. It is an architecture that is already partially implemented, in varying degrees of completeness, across the most sophisticated gesture systems currently deployed. The WRSLT system exhibits the second and fourth features explicitly, and approximates the third through semantic prototype alignment. Swipe keyboards implement the first, third, and fifth features in their most commercially mature form. Musical mimetic cognition, as described by Cox, operates through all five features simultaneously, which is why it serves as the theoretical reference point for what a fully developed embodied inference system should look like.

The broader significance of this convergence extends well beyond gesture recognition as a specialized engineering domain. Gesture recognition, understood as latent action reconstruction over admissibility manifolds, is a surface expression of a much more general problem: the problem of inferring embodied intention from observable behavior. That problem is central to human social cognition, developmental learning, clinical assessment, robotic learning from demonstration, AI safety through intent interpretation, and the design of interfaces between embodied human agents and computational systems. Progress in gesture recognition, properly understood, is progress toward a scientific account of embodied semantic agency.

This perspective also suggests a recalibration of research priorities. The field has historically been organized around benchmark performance on recognition accuracy across predefined vocabulary sets, a metric that rewards classifier performance on pre-segmented, seen-user data and systematically undervalues the capabilities that matter most for practical deployment: continuous decoding, user-independent generalization, sparse-sensing robustness, and real-time constraint integration. A reorientation toward the inverse reconstruction framework would focus attention on the properties of the learned admissibility landscape, the quality of inter-user gluing, the temporal coherence structure of the reconstruction process, and the efficiency of local adaptation mechanisms. These are harder to benchmark but more diagnostic of the system's fundamental capabilities.

The convergence toward semantic inverse architectures raises a deeper question: whether these organizational principles are merely contingent engineering solutions, or whether they reflect a more general strategy by which adaptive systems generate reliable structure under conditions of partial observability and combinatorial complexity. Recent work across developmental biology, morphogenesis, cortical evolution, and systems neuroscience suggests the latter.

Constraint Geometry, Active Media, and Biological Foundations of Admissibility

The preceding sections have argued that gesture recognition systems increasingly operate not through isolated symbolic classification but through reconstruction over constrained trajectory spaces. Sign language decoding, swipe typing, multimodal sensor fusion, and musical gesture analysis all converge toward the same operational structure: partial observations are integrated against learned priors over admissible embodied trajectories until a semantically stable interpretation emerges. The question that now arises is whether this convergence reflects merely a recurring engineering convenience, or whether it points toward a deeper organizational principle shared across biological and cognitive systems more generally.

Recent work in developmental biology, morphogenesis, cortical evolution, and systems neuroscience suggests the latter. Across these domains, an increasingly coherent picture is emerging in which biological complexity is generated not through exhaustive specification of final structure, but through dynamic manipulation of the constraints governing how signals propagate through evolving media. This framework — variously described in terms of active media, developmental accessibility landscapes, or constraint geometry — provides a powerful interpretive lens through which to understand not only biological development but the recent evolution of gesture recognition architectures.

The central critique motivating these approaches is fundamentally informational. The classical blueprint metaphor in biology assumes a direct mapping between genomic specification and final anatomical structure: genes are treated as architectural instructions specifying the positions and identities of components in the mature organism. This interpretation encounters a severe scaling problem when confronted with systems

such as the human neocortex, which contains on the order of sixteen billion neurons interconnected through hundreds of trillions of synaptic connections, while the information capacity of the genome is demonstrably insufficient to explicitly encode the final coordinates and connectivity structure of each component. The implication is not that biology compresses information efficiently, but that the explanatory ontology is incorrect. The genome does not specify final structure directly. Rather, it specifies evolving constraints governing which developmental trajectories remain accessible over time.

This shift in ontology maps precisely onto the framework developed throughout this review. The state manifold captures the full space of physically possible configurations available to the system. The admissibility structure is the dynamically evolving set of constraints determining which paths through that manifold are currently open. Stochastic fluctuations then drive exploration through admissible regions, producing the realized macroscopic structure at the observational level. Development becomes not the execution of a fixed plan, but controlled navigation through evolving accessibility landscapes. The blueprint metaphor is a snapshot model applied to a process that never stops moving.

The relevance of this framework to gesture recognition becomes vivid through the phenomenon of active media. In passive media, signals traverse substrates without altering the propagation geometry itself. In active media, every signal modifies the constraints governing future propagation. Biological development appears to be organized overwhelmingly through the latter principle. In vascular patterning studies of *Pilea peperomioides*, reticulate leaf vein networks were shown to closely approximate Voronoi tessellations generated around hydathode positions [25]. Rather than constructing veins through top-down specification, the developing leaf emits weak, unpolarized auxin diffusion waves from distributed sources; the collision boundaries between these waves stabilize into vascular structures. The global geometry emerges through local interference dynamics without centralized planning, and remains robust under environmental perturbation precisely because it is dynamically recomputed through local interactions rather than statically prescribed. A competing hierarchical partition model achieved a Jaccard overlap of approximately 0.40 against observed vein geometry, near the random-point baseline, while the hydathode-centered Voronoi model achieved approximately 0.72 — a gap decisive enough to distinguish the two organizational principles. What is biologically implemented is a distributed, constraint-based, locally-computed approx-

imation of an optimal spatial partition whose global coherence emerges entirely from local wave interference dynamics.

A more striking instance of reflexive medium modification appears in recent work on zebrafish blastoderm morphogenesis [22]. Nodal morphogen diffusion does not merely traverse a passive tissue substrate. Rather, the morphogen actively alters the mechanical properties of the tissue through Wnt11f2-mediated increases in cell adhesion, driving a rigidity phase transition that subsequently restricts further morphogen propagation. The signal modifies the medium through which future signals travel. The resulting developmental geometry is therefore reflexive in a strong sense: propagation dynamically alters the constraints governing subsequent propagation. What is especially significant is the causal direction. Optogenetic experiments demonstrated that artificially restoring the tissue's physical rigidity in mutant embryos unable to undergo the phase transition was sufficient to rescue normal morphogen gradient dynamics, without any direct manipulation of the chemical signaling pathways. The material geometry is causally upstream of the chemical signal. The geometry controls the chemistry.

This finding inverts the intuitive priority that most engineering approaches to biological mimicry have assumed. We tend to model biological systems as chemical signal processors in which geometry is a secondary consequence of signaling. What these experiments reveal is that biological systems are, fundamentally, geometrically organized active media in which chemical signals are secondary consequences of evolving physical accessibility structures.

The implications for cortical development reinforce this conclusion from a different direction. Comparative studies of human and murine cortical progenitors reveal that the dramatic expansion in human neocortical complexity relative to other mammals is not primarily attributable to the evolution of novel molecular machinery [23]. Human and mouse cortical progenitors share nearly identical genetic toolkits. The critical difference lies in temporal accessibility: human progenitors gain access to proliferative transcriptional programs earlier in development and sustain them for longer developmental windows. The gene JUNB, for example, activates in early radial glia in humans but only in post-mitotic mature neurons in mice. Because it activates early in the human developmental sequence, it drives prolonged proliferative expansion rather than merely modulating mature neuronal function. The latent computational program that generates neocortical complexity was present in the ancestral mammalian genome;

what changed was the accessibility geometry governing when that program became available.

Experimental confirmation of this principle is particularly striking. Artificially activating IRF1 — the upstream regulatory factor that physically unspools the chromatin around JUNB — early in mouse cortical progenitors caused those progenitors to exhibit human-like extended proliferative behavior. The mouse tissue did not lack the necessary molecular machinery. It lacked access to it at the critical developmental window. Evolution, in this account, acts not primarily by inventing new genomic content but by reorganizing the admissibility structure governing which content can be accessed at what time. The genome defines the state manifold; evolution reorganizes the constraint geometry that governs which trajectories through that manifold are developmentally accessible.

The parallel to gesture recognition is exact. Different users do not differ primarily in the semantic content available to them but in the admissibility geometry governing how that content is accessible from their motor trajectories. User-independent recognition requires learning a constraint structure that correctly captures which deformations of a gesture trajectory fall within the same semantic class and which fall outside it. The architecture must learn, in effect, the temporal accessibility structure of the gesture manifold — identifying which gesture variants are admissible realizations of a semantic target under individual motor variability, rather than treating all deviations from a mean trajectory as classification errors.

The deepest parallel between the biological framework and the gesture recognition framework appears in recent neuroscience of hippocampal memory routing. Studies of the hippocampal-retrosplenial communication circuit demonstrate that information routing in the brain depends not on static anatomical connectivity but on dynamically rotating communication subspaces [24]. The same physical pool of neurons in CA1 participates in different functional communication geometries depending on task context: the input subspace receiving from CA3 is nearly orthogonal to the output subspace projecting to retrosplenial cortex. Artificial networks constrained to maintain aligned input-output subspaces — even with full weight learning freedom — showed profoundly impaired routing performance, establishing that the orthogonal rotation is computationally load-bearing rather than incidental.

This result has direct implications for how we understand multimodal gesture recog-

nition. The effective computational geometry governing interpretation is not fixed by the physical sensor configuration or the trained network weights. It is dynamic, context-sensitive, and continuously reconfigurable through the same kind of subspace rotation that enables selective hippocampal routing. A mature gesture inference system may need to maintain multiple overlapping interpretive subspaces and dynamically rotate among them depending on linguistic context, user state, and task demands, rather than committing to a single fixed recognition geometry.

Sleep-state dynamics in the hippocampal circuit provide a further parallel to the temporal stabilization mechanisms discussed in the gesture recognition literature. During non-REM consolidation, the CA1-CA3 subspace — the internal hippocampal learning channel — remains highly plastic, continuously updating to encode new memories. But the CA1-retrosplenial subspace — the output channel to long-term cortical storage — remains stable, protecting previously consolidated content from interference. The system achieves simultaneous rapid learning and robust memory retention not through structural changes to the physical circuit, but by maintaining different geometric stability properties in different communication subspaces.

The gesture recognition analogue is the tension between generalization and personalization. A system that adapts too rapidly to individual motor variation risks overwriting its population-level admissibility model; one that adapts too slowly fails to capture individual-specific deviations. The biological solution — differential stability across communication subspaces, fast plasticity in one channel and slow stability in another — points toward an architectural solution that has not yet been fully explored in gesture recognition systems.

What emerges from this synthesis is a unified principle that connects developmental biology, cortical evolution, systems neuroscience, and gesture recognition under a single explanatory framework. Biological and cognitive systems alike generate reliable, complex, adaptive behavior not by explicitly specifying every outcome but by dynamically sculpting the geometry of the accessible. The signal modifies the medium. The medium constrains the signal. Complexity emerges from the iterated interaction between propagating signals and evolving accessibility structures, without any central authority maintaining a master plan.

Gesture recognition systems are, from this perspective, engineered approximations to the same organizational logic. They work best when they learn the admissibility

geometry of the motor manifold rather than merely fitting classifiers to observed trajectories. They fail characteristically when they treat the medium as passive — when they assume that observations are simple readouts of fixed underlying states rather than partial projections through continuously evolving constraint fields. The next generation of gesture systems will likely need to incorporate something closer to the active medium principle: architectures in which the system's prior model of admissible trajectories is itself continuously updated by the trajectories it observes, in a reflexive loop that mirrors the self-modifying geometry of biological development.

This is not merely a biological analogy. It is a claim about the structure of the problem itself. Embodied action unfolds in active media. Bodies modify their own motor manifolds through learning and habit. Social contexts modify the admissibility of gesture trajectories through convention and expectation. Physical environments modify the accessible action space through affordance and constraint. A gesture recognition system that models the medium as static is solving an easier but fundamentally different problem than the one actually posed by embodied human action. The convergence of gesture systems toward inverse reconstruction architectures is the field's practical discovery of this fact. The biological literature reviewed in this section suggests that evolution arrived at the same conclusion considerably earlier.

Open Problems and Future Directions

The theoretical framework developed here identifies several open problems that are likely to structure the next phase of progress in gesture recognition and embodied action understanding.

The most immediately pressing challenge is the development of richer admissibility landscape models that extend beyond the gesture vocabularies currently studied. The WRSLT system, to its credit, evaluates on a vocabulary of 200 words across two sign languages — substantially larger than most prior wireless systems — but still a small fraction of the functional vocabulary of fluent signers. Extending user-independent recognition to full natural sign language vocabularies will require admissibility models that capture the compositional structure of sign languages: the way in which signs are built from a finite set of phonological parameters (handshape, location, movement, and palm orientation) in ways that constrain which trajectories are linguistically well-formed and

which are not. The admissibility landscape at the sentence level is a dramatically more constrained object than the vocabulary-level landscape, and exploiting that structure is likely to be essential for practical continuous translation systems.

The cross-linguistic generalization problem is a further open challenge. The WRSLT system's evaluation on both ASL and ISL, with separately trained models, demonstrates that the architecture can accommodate multiple sign languages without fundamental modification. But the more interesting question is whether a single model can learn admissibility representations that generalize across sign languages, capturing the structural properties shared by sign languages as a family while adapting to the specific phonological conventions of each. The linguistic evidence suggests that sign languages share deep structural properties while differing in surface conventions, which in principle supports cross-linguistic transfer. Whether this transfer is achievable with current multimodal architectures remains an open empirical question [5].

The coarticulation problem at the trajectory level remains largely unsolved in practical systems. Current sliding-window approaches treat transitional gestures between signs as uninformative noise, correctly filtering them out through the repetition-based confirmation mechanism but not exploiting them for disambiguation or prediction. In fluent signers, however, transitional movements carry substantial information about the preceding and following signs, because the trajectory through the transition region is shaped by the coarticulation structure of the adjacent signs. A system capable of modeling this coarticulation structure could in principle use transitional movement patterns to update its predictions about the upcoming sign before the sign is fully executed, substantially improving both accuracy and latency.

The sparsity and miniaturization trajectory visible in the WRSLT architecture — from ten-finger full instrumentation to seven-finger sparse selection via LRP — points toward a further open question: how sparse can the sensing configuration become before performance degrades below practical thresholds? The theoretical framework suggests that the answer depends on the structure of the admissibility landscape and the complementarity of the retained projections. If the seven selected fingers provide near-maximally complementary constraints on the latent motor state, further reduction may be possible only if additional constraint sources are introduced — for example, language priors that compensate for reduced gestural resolution, or a small number of additional modalities (such as wrist-mounted IMUs for global arm position) that recover information lost by

removing peripheral finger sensors.

The music and expressive gesture frontier represents perhaps the most scientifically unexplored territory. The mimetic cognition framework suggests that human interpretation of expressive gesture operates through mechanisms substantially richer than current gesture recognition systems. Modeling the continuous expressive dimensions of gesture — the dimensions that carry affect, effort, and stylistic identity rather than propositional content — will require representations that extend beyond discrete semantic labels into continuous manifold geometry. This may require a fundamentally different evaluation paradigm, one organized around continuous similarity metrics and perceptual judgment rather than categorical accuracy.

Finally, the privacy and ethical dimensions of powerful gesture inference systems deserve explicit attention. A mature semantic inverse engine capable of reconstructing intended embodied actions from sparse, unobtrusive sensor data would be extraordinarily useful for accessibility technology, rehabilitation, human-computer interaction, and skill training. The same capability, deployed without appropriate constraint, would represent a significant extension of behavioral surveillance beyond what current vision or biometric systems provide. The architectural features that make gesture inference systems more useful — distributed sensing, personalized adaptation, continuous reconstruction, cross-context generalization — are also the features that make them most capable of reconstructing behavioral patterns that individuals may not intend to disclose. This tension is structural rather than incidental, and its resolution requires normative and regulatory frameworks that are currently substantially less developed than the technical capabilities they need to govern.

Conclusion

Gesture recognition has conventionally been treated as a classification problem, and the field has made substantial progress within that framing. But the persistent failure modes of classifier-based systems — degradation under continuous input, sensitivity to user-specific variation, fragility under real-world conditions, and the fundamental impossibility of resolving dynamical ambiguity from local observations — reveal that the framing is inadequate. The trajectory is the primary object, not the label. The label is an abstraction derived from the trajectory, not a property it possesses.

This review has argued that the field is converging toward a more adequate framing, one in which gesture systems are semantic inverse engines: systems that reconstruct latent embodied intention from partial, temporally extended, multimodal projections over constrained admissibility manifolds. This convergence is visible across engineering substrates as different as ring-type wireless wearables, swipe keyboard decoders, video-language models, and musical gesture analysis systems. It is also visible in the theoretical frameworks that humans use to interpret embodied action: Cox's mimetic hypothesis describes a cognitive architecture that already operates as a latent inverse inference engine, reconstructing intended embodied action from partial sensory traces through constraint-compatible simulation.

The key conceptual moves that define this emerging framework are four. First, the recognition of observable signals as projections rather than direct semantic content: the gesture is evidence, not the meaning. Second, the formulation of multimodal fusion as constraint multiplication over shared latent manifolds, rather than feature concatenation or channel voting. Third, the interpretation of user-independent generalization as a sheaf-theoretic gluing problem, in which the goal is constructing coherent global semantic sections from locally variable, signer-specific fiber data. Fourth, the characterization of temporal stabilization as constraint closure and irreversible semantic commitment, rather than probabilistic updating across independent time steps.

These moves collectively reframe gesture recognition as a special case of a much more general problem: the inverse problem of embodied semantic agency. Sign language, swipe typing, musical performance, and skilled tool use are not four separate gesture recognition problems. They are four surfaces of the same underlying challenge: inferring what an embodied agent intends to accomplish from the partial, noisy, temporally extended traces that its actions leave in the observable world.

The scientific frontier is not more accurate gesture classifiers. It is richer constraint models — systems that understand not just what a trajectory looks like, but what embodied world it makes sense inside. Progress toward that frontier will require theoretical frameworks adequate to the structure of the problem, empirical methods that evaluate the properties that matter for practical deployment, and hardware architectures that provide the complementary projections needed to make reconstruction tractable. The present moment, in which ring-type wireless wearables, multimodal foundation models, and embodied cognition theory are all simultaneously maturing, is an unusually

propitious one for work at this intersection.

Modern gesture systems are converging toward semantic inverse engines: systems that reconstruct latent embodied intention from partial, multimodal, temporally extended traces over constrained admissibility manifolds. That convergence is the field's most important current development, and understanding it as such is a prerequisite for directing the field's next phase of growth.

Appendices

Projection Operators and Admissibility Reconstruction

Let $\mathcal{M}$ denote the motor manifold and $C_t : \mathcal{M} \to \{0, 1\}$ the admissibility operator defined in Section 7, so that $\mathcal{A}_t = \{x \in \mathcal{M} \mid C_t(x) = 1\}$. Each sensing modality i provides a projection operator

$$\pi_i : \mathcal{M} \to \mathcal{O}_i,$$

mapping the latent motor state to an observation in modality i's observation space $\mathcal{O}_i$.

Definition (Residual uncertainty set). Given observations $o_1(t), \ldots, o_k(t)$ from k modalities at time t, the *residual uncertainty set* is

$$R_t = \bigcap_{i=1}^{k} \pi_i^{-1}\big(o_i(t)\big) \cap \mathcal{A}_t.$$

Proposition (Constraint multiplication). If the projections $\{\pi_i\}$ are sufficiently complementary — that is, if $\bigcap_i \ker(d\pi_i)$ has measure zero in $\mathcal{M}$ — then each additional modality j reduces the measure of R_t:

$$\mu(R_t^{(k+1)}) \leq \mu(R_t^{(k)}),$$

with strict inequality whenever π_{k+1} provides non-redundant constraint. Under sufficient complementarity, $\mu(R_t) \to 0$ as $k \to \infty$. Recognition is the identification of the

trajectory of maximum posterior weight $P(\gamma)$ within R_t.

Theorem (Projection Insufficiency). Let Γ_{coart} denote the space of coarticulated trajectories over $\mathcal{M}$ (as defined in Appendix B). For any finite family of local projection operators $\{\pi_i\}_{i=1}^{k}$ and any $\epsilon > 0$, there exist distinct trajectories $\gamma, \gamma' \in \Gamma_{\text{coart}}$ such that

$$\|\pi_i \circ \gamma(t) - \pi_i \circ \gamma'(t)\| < \epsilon \quad \text{for all } i \in \{1, \ldots, k\} \text{ and all } t \in [t_j, t_{j+1}],$$

yet γ and γ' belong to distinct semantic classes. That is, local projections are generically non-injective over the space of coarticulated trajectories: no finite collection of local modality projections suffices to disambiguate all semantically distinct coarticulated gesture streams.

Proof sketch. Coarticulation makes the semantic identity of a local segment dependent on its trajectory-level context (Appendix B). Two trajectories may agree arbitrarily closely at the local segment level while differing in adjacent segments in ways that change the local segment's semantic interpretation. Since local projections observe only the segment, not the full trajectory context, they cannot resolve this class of ambiguity. Disambiguation requires either temporal integration over a window spanning multiple adjacent segments or multimodal constraints that recover trajectory-level context from non-local observations.

Remark. This theorem unifies several major arguments in the paper simultaneously: the failure of isolated-segment classifiers, the necessity of temporal reconstruction, the importance of multimodal fusion, and the inevitability of probabilistic stabilization rather than instantaneous decoding. The WRSLT's sliding-window framework, swipe keyboard language priors, Cox's mimetic reconstruction, and the hippocampal subspace rotation mechanism are all engineering or biological responses to the same structural insufficiency.

Remark. The WRSLT system's seven-finger sparse sensing configuration corresponds to a specific choice of $k = 7$ complementary projections. The layer-wise relevance propagation finger selection procedure identifies the subset of available projections that reduces $\mu(R_t)$ most efficiently, operationalizing the complementarity condition above.

Coarticulation as Trajectory Coupling

Let gestures be modeled as continuous curves $\gamma : [0, T] \to \mathcal{M}$ through the motor manifold. In a discrete-token model, the trajectory is piecewise constant in semantic identity across non-overlapping intervals:

$$\gamma(t) = \gamma_j(t) \quad \text{for } t \in [t_j, t_{j+1}],$$

where each segment γ_j is determined independently by a local semantic target and the segments share only endpoint continuity $\gamma_j(t_{j+1}) = \gamma_{j+1}(t_{j+1})$. The isolated-classification paradigm is an instance of this model: each γ_j is treated as if it were generated by its local target alone.

Definition (Coarticulated trajectory). A trajectory γ is *coarticulated* if each local segment satisfies

$$\gamma_j(t) = F_j\big(\gamma_{j-1},\, \gamma_{j+1},\, C_t\big) \quad \text{for } t \in [t_j, t_{j+1}],$$

where C_t is the active admissibility operator at time t and F_j is a functional depending on adjacent segments as well as the local semantic target. The segment γ_j is not determined by its target alone; it is continuously deformed by anticipatory preparation for γ_{j+1} and residual dynamics from γ_{j-1}. The space of coarticulated trajectories over $\mathcal{M}$ is denoted Γ_{coart}.

Generative mechanism: hierarchical oscillator dynamics. The functional F_j is not merely a formal abstraction; it has a concrete mechanistic interpretation in terms of hierarchical coupled oscillatory dynamics. Define a network of n motor oscillatory subsystems, each characterized by a phase state $\theta_i(t) \in [0, 2\pi)$. The coupled phase dynamics are governed by:

$$\dot{\theta}_i = \omega_i + \sum_j K_{ij} \sin(\theta_j - \theta_i) + I_i(t),$$

where ω_i is the intrinsic oscillatory tendency of subsystem i, K_{ij} is the coupling strength between subsystems i and j, and $I_i(t)$ is a higher-order semantic modulation signal encoding the current gestural intention. Observable gesture trajectories emerge

as smooth projections from the latent oscillatory coordination state:

$$\gamma(t) = \Pi\big(\theta_1(t), \ldots, \theta_n(t)\big),$$

where Π maps latent oscillator phases into observable motor geometry in $\mathcal{M}$. The coupling matrix $K = (K_{ij})$ encodes the hierarchical structure of the motor coordination system: strong coupling within functional subsystems (e.g., the fingers of one hand), weaker coupling across subsystems, and asymmetric coupling reflecting the temporal organization of gesture sequences.

Semantic intention operates through $I_i(t)$, which biases the coupled system toward phase-locked configurations corresponding to the target gesture class, rather than specifying the trajectory pointwise. Different stable phase-locked configurations of the oscillator network correspond to different gesture attractor basins in $\mathcal{M}$. The admissibility landscape $\mathcal{A}_t$ is precisely the image under Π of the set of dynamically stable phase-locked configurations available to the system at time t:

$$\mathcal{A}_t = \Pi\big(\{\theta \in T^n \mid \theta \text{ is a stable equilibrium of the coupled system at time } t\}\big).$$

Coarticulation as entrainment. Under the oscillator framework, coarticulation is not an imperfection but a structural consequence of coupling dynamics. The phase state $\theta_i(t)$ for subsystems involved in gesture γ_{j+1} begins to evolve toward the target configuration before the current gesture γ_j completes, because coupling propagates continuously through the oscillator network. Residual phase coherence from γ_{j-1} likewise persists beyond its nominal completion. The coupling functional F_j therefore encodes the anticipatory and residual phase dynamics generated by the oscillator network, rather than an abstract trajectory coupling. Gesture boundaries become soft synchronization transitions — phase-locking events — rather than discrete symbolic separations. This provides a mechanistic foundation for the argument of Section 4 that segmentation is not ontologically primary.

Individual variation as oscillator geometry. Different users correspond to different coupling matrices K and intrinsic frequencies ω_i, reflecting individual differences in anatomy, motor history, and learned coordination patterns. Signers who share the same seman-

tic gesture target (the same stable phase-locking configuration in the abstract oscillator sense) will nonetheless generate different trajectories $\gamma(t) = \Pi(\theta(t))$ if their coupling geometries differ. This is exactly the structure of local sections in the sheaf-theoretic framework of Appendix C: different users realize the same semantic class through different oscillator trajectories, compatible in semantic content but varying in motor geometry. The obstruction class $[\omega] \in \check{H}^1(\mathcal{U}, \mathcal{F})$ can now be interpreted concretely as the measure of incompatibility between users' oscillator phase-space foliations when projected onto the shared semantic manifold.

Active coupling geometry. The coupling matrix itself evolves through motor learning, fatigue, adaptation, and contextual priming:

$$\frac{\partial K_{ij}}{\partial t} = \Phi(S_t, K_{ij}),$$

where S_t represents ongoing sensory and semantic modulation. This makes the motor coordination substrate itself an active medium in the sense of Appendix D: each gesture execution modifies the coupling geometry through which future gestures will be generated. Motor learning, habit formation, expressive style, and the long-term consequences of practice all correspond to evolution of K under repeated trajectory execution. The motor system is not a fixed transducer converting semantic intention to motion; it is an evolving dynamical medium whose generative geometry changes with use.

Proposition (Ontological instability of local classification). If $\gamma \in \Gamma_{\text{coart}}$, then for any fixed classifier $f : \mathcal{O}_i \to L$ over a label set L, there exist trajectories $\gamma, \gamma' \in \Gamma_{\text{coart}}$ such that $\pi_i \circ \gamma_j = \pi_i \circ \gamma'_j$ on $[t_j, t_{j+1}]$ (identical local projection at segment j) while the correct label of γ_j differs from the correct label of γ'_j (different semantic interpretation arising from different trajectory-level context). Local classifiers are therefore structurally incomplete for coarticulated inputs regardless of their capacity. The error is not quantitative but ontological: the representational primitive (the local segment projection) does not contain sufficient information to determine the target quantity (the segment's semantic role within the full trajectory).

Remark. Coarticulation introduces dynamical ambiguity distinct from geometric and semantic ambiguity. Two trajectory segments may have identical local geometry and identical semantic target yet belong to different dynamical regimes depending on their temporal neighbors. This form of ambiguity is invisible to any local-window classifier

and requires trajectory-level temporal integration for resolution.

Remark. The sliding-window confirmation mechanism in the WRSLT framework is a minimal response to coarticulation: it delays commitment until the trajectory has persisted across overlapping windows, partially recovering the trajectory-level context that purely local classifiers discard. The full resolution of coarticulation would require explicitly modeling the coupling functional F_j, exploiting transitional gesture dynamics as predictive signals for adjacent segment identification.

User-Independent Generalization as Sheaf-Theoretic Gluing

Let $\mathcal{G}$ denote the space of gesture classes (a finite set in practice) and let $\mathcal{U} = \{U_\alpha\}$ be an open cover of the space of users, where each U_α represents a local neighborhood of users with sufficiently similar motor geometry. For each U_α, let $\mathcal{F}(U_\alpha)$ denote the set of local gesture-semantics correspondences: functions $s_\alpha : \mathcal{T}_{U_\alpha} \to \mathcal{G}$ mapping motor trajectories observed from users in U_α to gesture classes.

Definition (Local section). A *local section* $s_\alpha \in \mathcal{F}(U_\alpha)$ is a locally consistent gesture-semantics correspondence: a function that correctly classifies gestures for users within U_α under the motor geometry characteristic of that neighborhood.

Definition (Gluing condition). Local sections s_α and s_β are *compatible* on $U_\alpha \cap U_\beta$ if

$$s_\alpha\big|_{U_\alpha \cap U_\beta} = s_\beta\big|_{U_\alpha \cap U_\beta}.$$

A *global section* $s \in \mathcal{F}(\bigcup_\alpha U_\alpha)$ exists if and only if all pairs of local sections are compatible on their overlaps.

Proposition (Obstruction to user-independent recognition). The obstruction to constructing a coherent global section from locally consistent signer-specific sections is measured by the Čech cohomology class

$$[\omega] \in \check{H}^1(\mathcal{U}, \mathcal{F}).$$

User-independent recognition is achievable if and only if $[\omega] = 0$, i.e., all local sections can be coherently glued into a global section without contradiction. When $[\omega] \neq 0$,

no single classifier generalizes across the full user population, and per-user adaptation is structurally necessary rather than merely convenient.

Remark. The WRSLT system's semantic prototype alignment — projecting sensor embeddings into a shared latent space organized by text embeddings of word meanings — is an engineering attempt to construct a representation in which the gluing condition is approximately satisfied. The residual unseen-user error (approximately 11–12% at the 100-word scale) reflects the degree to which the obstruction class $[\omega]$ remains nontrivial in the learned embedding geometry. Personalized user dictionaries in swipe keyboard systems correspond to the construction of locally refined sections s_α, deferring the global gluing problem to incremental online adaptation.

Active Media and Reflexive Propagation Kernels

Definition (Propagation kernel). A *propagation kernel* $K_t(x, y)$ describes how a signal, developmental program, or information pattern at state y influences the future state at x given the current geometry of the medium at time t:

$$S(x, t + \delta t) = \int K_t(x, y)\, S(y, t)\, dy + \text{source terms}.$$

Definition (Passive medium). A medium is *passive* if the propagation kernel is invariant under signal passage:

$$\frac{\partial K_t}{\partial S} = 0.$$

In a passive medium, signals traverse the substrate without altering the geometry governing subsequent propagation. Standard engineering communication channels (wires, optical fibers, wireless channels on fixed substrates) approximate this condition.

Definition (Active medium). A medium is *active* if

$$\frac{\partial K_t}{\partial S} \neq 0,$$

meaning the propagation kernel evolves as a functional of the signal field passing through it. The medium participates in the dynamics it hosts rather than merely trans-

mitting them. The coupled dynamics are

$$\frac{\partial K_t}{\partial t} = \Phi(S_t, K_t),$$

where Φ captures the mutual regulation of signal and geometry. Development, motor learning, and neural plasticity are all instances of active-medium dynamics: each signal passage reshapes the kernel governing future propagation.

Proposition (Biological systems as active media). Let S denote a biological signaling field (auxin concentration, morphogen gradient, transcription factor activity, or neural population activity) and K_t the associated propagation kernel. In each of the four biological systems examined in Section 9 of this paper, $\partial K_t/\partial S \neq 0$: the signal modifies the medium through which future signals propagate. Consequently, biological development and cognition cannot be modeled as propagation through a fixed passive medium; the kernel itself is the primary object of regulatory control.

Remark. Gesture recognition systems that employ fixed trained classifiers are passive-medium approximations: the network weights (the kernel) are fixed after training and do not adapt to the gesture stream being processed. Systems employing online adaptation, personalized decoding, or predictive updating of prior models are approximations to active-medium operation. The convergence of gesture systems toward adaptive, personalized, and temporally evolving architectures corresponds to the discovery that embodied motor systems are active media requiring active-medium inference engines.

Remark (Motor systems as active oscillatory media). The hierarchical oscillator coupling framework of Appendix B provides a concrete biological instantiation of the active-medium principle. The coupling matrix K_{ij} governing motor coordination corresponds to the propagation kernel $K_t(x, y)$ in the active-media framework. The evolution equation $\partial_t K_{ij} = \Phi(S_t, K_{ij})$ is a specific instance of the general active-medium condition $\partial K_t/\partial S \neq 0$: motor coordination geometry evolves under the influence of the gesture trajectories it generates, through mechanisms of synaptic plasticity, motor adaptation, and habitual entrainment. A gesture recognition system capable of tracking this evolution — updating its model of the user's coupling geometry as new gestures are observed — would constitute a fully active-medium inference engine operating over embodied motor dynamics.

Temporal Stabilization and Irreversible Semantic Commitment

Definition (Trajectory ensemble entropy). At time t, let $\Gamma_t = \{\gamma \in \Gamma_{\text{coart}} \mid \gamma \text{ consistent with evidence up to } t\}$ denote the admissible trajectory ensemble given accumulated observations. The *trajectory ensemble entropy* is

$$H_t = -\int_{\Gamma_t} P(\gamma) \log P(\gamma)\, d\gamma,$$

where $P(\gamma) \propto e^{-\mathcal{S}[\gamma]}$ is the trajectory measure defined in Section 7. H_t measures the remaining uncertainty over admissible trajectory interpretations given all evidence accumulated to time t. When H_t is large, many competing trajectory hypotheses remain plausible. When H_t is small, the evidence has concentrated the posterior near a single trajectory or a narrow equivalence class of trajectories.

Definition (Semantic stabilization). A trajectory $\gamma|_{[0,t]}$ achieves *semantic stabilization* at time t^* if

$$\left.\frac{dH_t}{dt}\right|_{t=t^*} \to 0 \quad \text{and} \quad H_{t^*} \leq H^*,$$

where H^* is a commitment threshold below which the trajectory ensemble is sufficiently concentrated on a single semantic class to justify irreversible commitment.

Definition (Irreversible commitment event). The *commitment event* $\Omega_{t^*} \to \ell^*$ is the transition from open hypothesis state to committed interpretation $\ell^* = \arg\max_\ell \int_{\gamma \in \ell} P(\gamma)\, d\gamma$, constituting an irreversible collapse of the admissible trajectory ensemble to the highest-weight semantic class. This maps to the KES framework's $\Omega_t \to H_{t+1}$ transition: the evolving admissible ensemble Ω_t collapses to a committed historical record H_{t+1} at the moment of irreversible semantic stabilization.

Proposition (Constraint closure as entropy reduction). Each overlapping window observation $o(t_k)$ eliminates trajectories inconsistent with the observed segment, reducing the measure of Γ_t. Under the trajectory measure P, this induces a monotonically non-increasing sequence

$$H_{t_1} \geq H_{t_2} \geq \cdots \geq H_{t_n},$$

with strict inequality whenever the observation provides non-redundant constraint. Temporal stabilization is therefore equivalent to constraint closure over the admissible trajectory ensemble: the successive elimination of incompatible hypotheses under accumulating evidence, converging toward commitment at the threshold H^*.

Remark. In the WRSLT sliding-window framework, the commitment threshold is implemented as a count criterion: a gesture label is committed when it has appeared in two or more consecutive overlapping windows. This is a discrete approximation to the continuous entropy criterion above. The window count serves as a proxy for $H_t < H^*$: sufficient consecutive appearances imply that the trajectory posterior has concentrated around a single class. The irreversibility of commitment corresponds to the treatment of confirmed words as fixed elements of the emerging sentence, not subject to revision absent explicit correction — an architectural choice that reflects the broader principle that semantic commitment in active-medium inference systems is structurally irreversible once constraint closure is achieved.

Semantic Accessibility Curvature

The admissibility landscape $\mathcal{A}_t \subseteq \mathcal{M}$ has internal geometric structure beyond the bare question of which states are admissible. Some regions of $\mathcal{A}_t$ are densely accessible — many nearby trajectories in $\mathcal{M}$ are also in $\mathcal{A}_t$, so small perturbations remain semantically interpretable. Other regions are sparsely accessible — the admissible region is thin, and small perturbations may exit $\mathcal{A}_t$ entirely, producing classification failure or ambiguity.

Definition (Accessibility density). Let $\rho_{\mathcal{A}}(x)$ denote a smooth density function on $\mathcal{M}$ encoding the local measure of admissibility: high where many nearby trajectories belong to $\mathcal{A}_t$, low near the boundary of the admissible region. Define the *semantic accessibility curvature* at $x \in \mathcal{A}_t$ as

$$\kappa_{\mathcal{A}}(x) = \nabla^2 \log \rho_{\mathcal{A}}(x).$$

Proposition (Curvature and classifier margin). Regions of positive curvature ($\kappa_{\mathcal{A}}(x) > 0$) correspond to deep attractor basins of the admissibility landscape: perturbations to a trajectory in such a region tend to remain within $\mathcal{A}_t$, and any classifier trained over this region will find a large margin separating it from adjacent semantic classes. Regions of

negative curvature ($\kappa_{\mathcal{A}}(x) < 0$) correspond to boundary or saddle regions: perturbations are likely to exit the admissible class, and classifiers will exhibit small or unstable decision boundaries there. Formally, for a linear classifier with margin $\delta(x)$ over $\mathcal{M}$, we have the asymptotic relationship

$$\delta(x) \sim \rho_{\mathcal{A}}(x)^{1/2} \cdot \kappa_{\mathcal{A}}(x)^{-1/2}$$

in regions where the admissibility landscape is locally approximately Gaussian, connecting accessibility curvature directly to the geometric margin of the recognition boundary.

Remark. The distinction between "easy" and "hard" gesture classes in recognition systems is therefore not merely a function of visual or inertial complexity, but of the curvature structure of the semantic accessibility landscape. Gesture classes whose realization across users occupies high-curvature regions of $\mathcal{A}_t$ will be robust to inter-user variation and trainable in a user-independent setting. Classes realized near low-curvature boundaries require either more training data, more precise sensing, or user-specific calibration. This gives a geometric characterization of when personalization is necessary: not when individual users are unusual, but when the relevant gesture class occupies a low-curvature region of the shared admissibility landscape.

Remark (Connection to sheaf gluing). Accessibility curvature also provides a natural measure of the difficulty of gluing local sections in the framework of Appendix C. Users whose motor habits place the same gesture class in high-curvature regions of $\mathcal{A}_t$ produce local sections that are easy to glue into a global section, because the large-margin geometry provides ample overlap between signer neighborhoods. Users whose habits place the class near boundary regions with low or negative curvature introduce obstruction to global section construction: the local sections are locally consistent but globally incompatible because the margin is too small to accommodate inter-user variation. Personalization corresponds to locally deepening the curvature of $\rho_{\mathcal{A}}$ in the neighborhood of an individual user's motor geometry, enlarging the local attractor basin and increasing the effective margin, thereby reducing the obstruction class $[\omega]$ toward zero in the learned representation.

References

[1] J. Park, Y. Shin, I. S. Min, Y. Lee, J. Lee, J.-H. Hong, S. Park, S. H. Park, J. Baek, T. Kim, K. Kim, D. Kim, Y. Park, J. Kim, Y. U. Cho, Y. Choi, B. Jeon, K. Kang, D. Hwang, and K. J. Yu, "An AI-driven, wearable, conformal ring system for real-time and user-independent sign language interpretation," *Science Advances*, vol. 12, p. eaec8995, May 2026.

[2] S. Sharma and S. Singh, "Vision-based hand gesture recognition using deep learning for the interpretation of sign language," *Expert Systems with Applications*, vol. 182, p. 115657, 2021.

[3] D. Sarma and M. K. Bhuyan, "Methods, databases and recent advancement of vision-based hand gesture recognition for HCI systems: A review," *SN Computer Science*, vol. 2, p. 436, 2021.

[4] M. M. Zaki and S. I. Shaheen, "Sign language recognition using a combination of new vision-based features," *Pattern Recognition Letters*, vol. 32, pp. 572–577, 2011.

[5] K. Kudrinko, E. Flavin, X. Zhu, and Q. Li, "Wearable sensor-based sign language recognition: A comprehensive review," *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 82–97, 2021.

[6] A. Cox, *Music and Embodied Cognition: Listening, Moving, Feeling, and Thinking*. Bloomington: Indiana University Press, 2016.

[7] B. G. Lee and S. M. Lee, "Smart wearable hand device for sign language interpretation system with sensors fusion," *IEEE Sensors Journal*, vol. 18, pp. 1224–1232, 2017.

[8] N. M. Kakoty and M. D. Sharma, "Recognition of sign language alphabets and numbers based on hand kinematics using a data glove," *Procedia Computer Science*, vol. 133, pp. 55–62, 2018.

[9] J. Wu, L. Sun, and R. Jafari, "A wearable system for recognizing American sign language in real-time using IMU and surface EMG sensors," *IEEE Journal of Biomedical and Health Informatics*, vol. 20, pp. 1281–1290, 2016.

[10] J. Wu, Z. Tian, L. Sun, L. Estevez, and R. Jafari, "Real-time American Sign Language recognition using wrist-worn motion and surface EMG sensors," in *Proc. IEEE 12th*

Int. Conf. Wearable and Implantable Body Sensor Networks (BSN), 2015, pp. 1–6.

[11] Q. Zhang, J. Jing, D. Wang, and R. Zhao, "WearSign: Pushing the limit of sign language translation using inertial and EMG wearables," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, pp. 1–27, 2022.

[12] F. Wen, Z. Zhang, T. He, and C. Lee, "AI enabled sign language recognition and VR space bidirectional communication using triboelectric smart glove," *Nature Communications*, vol. 12, p. 5378, 2021.

[13] Z. Zhou, K. Chen, X. Li, S. Zhang, Y. Wu, Y. Zhou, K. Meng, C. Sun, Q. He, and W. Fan, "Sign-to-speech translation using machine-learning-assisted stretchable sensor arrays," *Nature Electronics*, vol. 3, pp. 571–578, 2020.

[14] A. Tashakori, Z. Jiang, A. Servati, S. Soltanian, H. Narayana, K. Le, C. Nakayama, C.-l. Yang, Z. J. Wang, J. J. Eng, and P. Servati, "Capturing complex hand movements and object interactions using machine learning-powered stretchable smart textile gloves," *Nature Machine Intelligence*, vol. 6, pp. 106–118, 2024.

[15] K. R. Pyun, K. Kwon, M. J. Yoo, K. K. Kim, D. Gong, W.-H. Yeo, S. Han, and S. H. Ko, "Machine-learned wearable sensors for real-time hand-motion recognition: toward practical applications," *National Science Review*, vol. 11, p. nwad298, 2024.

[16] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, and J. Clark, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Machine Learning (ICML)*, 2021, pp. 8748–8763.

[17] A. Binder, G. Montavon, S. Lapuschkin, K.-R. Müller, and W. Samek, "Layer-wise relevance propagation for neural networks with local renormalization layers," in *Artificial Neural Networks and Machine Learning – ICANN 2016*, Springer, 2016, pp. 63–71.

[18] Z. Sun, M. Zhu, X. Shan, and C. Lee, "Augmented tactile-perception and haptic-feedback rings as human-machine interfaces aiming for immersive interactions," *Nature Communications*, vol. 13, p. 5224, 2022.

[19] Y. Liu, X. Jiang, X. Yu, H. Ye, C. Ma, W. Wang, and Y. Hu, "A wearable system for sign language recognition enabled by a convolutional neural network," *Nano Energy*, vol. 116, p. 108767, 2023.

[20] X. Wu, X. Luo, Z. Song, Y. Bai, B. Zhang, and G. Zhang, "Ultra-robust and sensitive flexible strain sensor for real-time and wearable sign language translation," *Advanced Functional Materials*, vol. 33, p. 2303504, 2023.

[21] M. M. Nomeland and R. E. Nomeland, *The Deaf Community in America: History in the Making*. Jefferson, NC: McFarland, 2011.

[22] C. Autorino, D. Khoromskaia, L. Harari, E. Floris, H. Booth, C. Pallares-Cartes, V. Petrasiunaite, M. Dorrity, B. Corominas-Murtra, Z. Hadjivasiliou, and N. I. Petridou, "Tissue rigidity phase transition shapes morphogen gradients," *Nature Cell Biology*, 2026. doi:10.1038/s41556-026-01954-4.

[23] A. Javed et al., "Developmental gene expression patterns driving species-specific cortical features," *Nature*, 2026. doi:10.1038/s41586-026-10491-x.

[24] J. Gonzalez, M. Vöröslakos, D. Aykan, N. Soto, N. Nitzan, R. Swanson, M. Karadas, Z. S. Chen, and G. Buzsáki, "Subspace communication in the hippocampal–retrosplenial axis," *Nature*, 2026. doi:10.1038/s41586-026-10481-z.

[25] C. X. Zheng, S. Palit, M. Venezia, E. Blum, U. V. Pedmale, D. Jackson, E. Scarpella, P. Prusinkiewicz, and S. Navlakha, "Reticulate leaf venation in *Pilea peperomioides* is a Voronoi diagram," *Nature Communications*, vol. 17, p. 4111, 2026. doi:10.1038/s41467-026-71768-3.