A Layout-Free Text Input Topology via Dual Variational Autoencoders and Linear-Complexity Temporal Path Integration

Andrej Tetkić
Email: andrej@tetkic.com

***Abstract*—Conventional text entry interfaces remain fundamentally constrained by the legacy of mechanical typewriters, relying on static physical or virtual QWERTY layouts that disregard individual anatomical differences and user motor variability. This paper presents a machine learning framework that reformulates text entry as an online geometric regression problem between two high-dimensional manifolds, translating free-form multi-finger gestures directly into text sequences. The proposed architecture processes touch-pressure frames through a dual Variational Autoencoder (VAE) pipeline: a convolutional Spatio-Geometric Autoencoder compresses spatial contact topologies into a continuous latent space, while a recurrent Linguistic Sequence Autoencoder defines a stable linguistic target space. To map the temporal trajectory of gestures, we introduce a lightweight path-integral temporal transformation that summarizes variable-length sequences with a strict linear complexity of $O(N \cdot D)$. Cross-space alignment is achieved via a non-parametric Inverse Distance Weighting (IDW) interpolation scheme augmented with local spatial offsets to construct a self-correcting, "magnetic" mapping that eliminates catastrophic forgetting. Empirical validation yields a 94.46% pixel reconstruction accuracy for the spatial encoder, over 99% sequence reconstruction accuracy for short-to-medium n-grams within the textual decoder, and a 98% classification accuracy in one-shot mathematical validation trials. This paper serves as a theoretical and algorithmic foundation, evaluating the mathematical viability and spatial-temporal compression of the architecture via synthetic and offline benchmarks, establishing the groundwork for future live human-subject deployment.**

***Index Terms*—Representation learning, continuous manifold alignment, non-parametric online learning, geometric deep learning, time-series integration.**

I. Introduction

Standard text input interfaces structurally constrain human motor behavior to rigid, discrete coordinate mapping tasks. While single-finger continuous tracking systems like SHARK [1] or Swype exist, they fundamentally rely on absolute visual backdrops and struggle to parse the high-dimensional complexity of free-form, multi-contact gestures. To bypass these limitations, this work reformulates text entry not as a Human-Computer Interaction (HCI) interface challenge, but as an online geometric regression and continuous modality translation problem.

Inspired by cross-modal representation learning frameworks such as CLIP [2], we propose an architecture designed to map unsegmented spatiotemporal trajectories directly into a continuous linguistic manifold. Utilizing multi-finger text generation as our primary evaluative domain, the system projects these trajectories using frozen dual latent spaces. These stable manifolds are bridged by a dynamic, non-parametric spatial database, which elegantly decouples fundamental representation learning from user-specific optimization. This approach allows the system to achieve one-shot adaptation to a user's unique biomechanical tendencies without suffering from the catastrophic forgetting typical of parametric online fine-tuning.

This paper focuses strictly on the mathematical formulation, architectural design, and machine learning principles underlying this framework, serving as an algorithmic foundation prior to future human-in-the-loop HCI deployment. Specifically, our core algorithmic contributions are:

- **Linear-Complexity Temporal Path Integration:** A novel sequence summarization operator utilizing continuous path integrals that achieves strict $O(N \cdot D)$ complexity. This avoids the exponential feature explosion of deep signature transforms while remaining invariant to sensor sampling density and execution speed.
- **Dual-Manifold Non-Parametric Alignment:** A frozen dual-Variational Autoencoder (VAE) architecture bridged by an offset-stabilized Inverse Distance Weighting (IDW) interpolation scheme, enabling stable, one-shot online adaptation without catastrophic forgetting.
- **Invariant Spatial Representation:** A dynamic 360-degree polar projection preprocessing pipeline that ensures absolute positional invariance, forcing the network to extract generalized geometric topologies from raw multi-contact sensor frames.

II. Signal Preprocessing and Spatial Invariance

The theoretical input to the system consists of a continuous sequence of two-dimensional matrices $R_t \in \mathbb{R}^{64\times64}$ representing a monochromatic touch-pressure frame at time t. To maximize computational efficiency on general-purpose processors, downstream modeling is gated by a change-detection mechanism designated as Flux (Φ_t). Flux computes the sum of absolute differences between consecutive sensor frames.

Neural processing is triggered exclusively when $\Phi_t > \theta_{\text{flux}}$. Upon activation, the frame's Center of Gravity (CoG) is calculated to establish a spatial anchor point. To suppress high-frequency sensor noise and unintentional micro-tremors, the raw CoG coordinates are smoothed using a low-pass filter.

To ensure the system remains completely invariant to absolute hand placement on a sensing surface, the normalized

frame is transformed into a 360-degree polar projection centered around the filtered CoG coordinate. This transformation maps absolute coordinate profiles into position-invariant relative geometries, allowing the network to parse the structural shape of a hand posture regardless of where the interaction occurs. The resulting invariant tensor serves as the primary input for the dual-manifold neural pipeline.

III. Dual Latent Space Architecture

Traditional gesture recognizers reduce text entry to discrete classification or direct string mapping via simple nearest-neighbor algorithms. These methodologies treat the output space as a set of disjoint symbols lacking internal topology, which results in binary error rates and a rigid interface that cannot gracefully interpolate between known states.

In contrast, our framework structures the translation task as an online geometric regression between two continuous, high-dimensional manifolds. While mapping natural language to continuous latent spaces has been explored at the sentence level [3], our architecture targets sub-word character sequences (n-grams).

By employing a pre-trained and frozen character-based Linguistic Sequence Autoencoder, we establish a geometrically stable linguistic manifold where morphologically similar sequences (e.g., "mi" and "mo") reside at proportional Euclidean distances. This geometric structure allows the system to perform structured confidence estimation and exhibit smooth performance degradation under perceptual drift, rather than catastrophic failure. Crucially, freezing the weights of both the spatial and linguistic autoencoders eliminates the problem of catastrophic forgetting during user optimization; continuous adaptation is achieved entirely by dynamically updating non-parametric anchor mappings between the two stable spaces.

The overall end-to-end data progression through these modular components is illustrated in Fig. 1.

A. Spatio-Geometric Autoencoder (Spatial Convolutional Encoder)

Spatial layout configurations are processed and compressed using a lightweight, convolutional Variational Autoencoder (CNN-VAE) [4] optimized for low-latency CPU execution. The encoder maps each position-invariant polar touch frame into a 128-dimensional continuous vector space ($D = 128$). To force the network to extract robust, generalized structural features of contact profiles, the model was trained via unsupervised learning on a synthetic dataset containing 100,000 images representing random anatomical and non-anatomical multi-touch configurations.

B. Linguistic Sequence Autoencoder

The target language manifold is formed using a recurrent VAE built upon Gated Recurrent Unit (GRU) cells. This sequence-to-sequence model embeds variable-length n-grams (spanning 1 to 30 characters from the Latin alphabet) into a co-dimensioned 128D latent space. To guarantee immediate feedback during text composition, the recurrent decoder configuration balances parameter density and execution speed to ensure rapid inference on standard CPUs.

IV. Linear-Complexity Temporal Path Integration

The continuous stream of embeddings produced by the Spatio-Geometric Autoencoder forms a variable-length trajectory matrix of shape $(t \times D)$. Because real-time human gestures cannot be segmented explicitly prior to interpretation, standard temporal architectures (such as Recurrent Neural Networks or Deep Temporal Convolutional Networks) introduce systematic latency and obscure geometric transparency. Alternatively, mathematical path-encoding options like the truncated Signature Method [5], [6] scale geometrically with depth ($O(N \cdot D^k)$), introducing severe computational bottlenecks when evaluating multi-window sequences.

To achieve scale-invariance, direction-dependence, and high execution efficiency, we implement a custom, lightweight temporal path transformation of strict linear complexity.

A. Mathematical Formulation

Let an ordered discrete sequence of N latent spatial embeddings be defined as $P = \{p_1, p_2, \ldots, p_N\}$, where $p_i \in \mathbb{R}^D$. We treat this sequence as a continuous, piecewise linear curve $p(s)$ parameterized by its absolute cumulative arc length $s \in [0, L]$. The total geometric path length L is computed as:

$$L = \sum_{i=1}^{N-1} \|p_{i+1} - p_i\|_2 \tag{1}$$

To eliminate operational variances caused by the velocity of user execution, the trajectory path is reparameterized using a normalized arc-length variable $t = s/L$, bounding the domain to $t \in [0, 1]$. The temporal transformation encodes the trajectory into a fixed-dimensional summary vector $z \in \mathbb{R}^D$ by evaluating an element-wise line integral modulated by a static harmonic basis:

$$z = \int_0^1 p(t) \odot \cos(\omega t + \phi) dt \tag{2}$$

where $\odot$ denotes the Hadamard product. The frequencies $\omega \in \mathbb{R}^D$ are sampled from a log-uniform distribution $\omega_d \sim \exp(\mathcal{U}(\ln(\omega_{\min}), \ln(\alpha \cdot \omega_{\max})))$, where α governs high-frequency sensitivity. The phase shifts $\phi \in [0, 2\pi]^D$ are uniformly distributed to preserve dimensional orthogonality.

In practice, this continuous integral is solved directly via a discrete midpoint rule numerical integration across the path segments. For each individual segment i bounded by (p_i, p_{i+1}), its localized Euclidean length is $l_i = \|p_{i+1} - p_i\|_2$, yielding a fractional time delta $\Delta t_i = l_i / L$. The spatial midpoint $\bar{p}_i$ and its corresponding normalized temporal midpoint $\bar{t}_i$ are given by:

$$\bar{p}_i = \frac{p_i + p_{i+1}}{2}, \quad \bar{t}_i = \frac{1}{L} \left(\sum_{j=1}^{i-1} l_j + \frac{l_i}{2} \right) \tag{3}$$

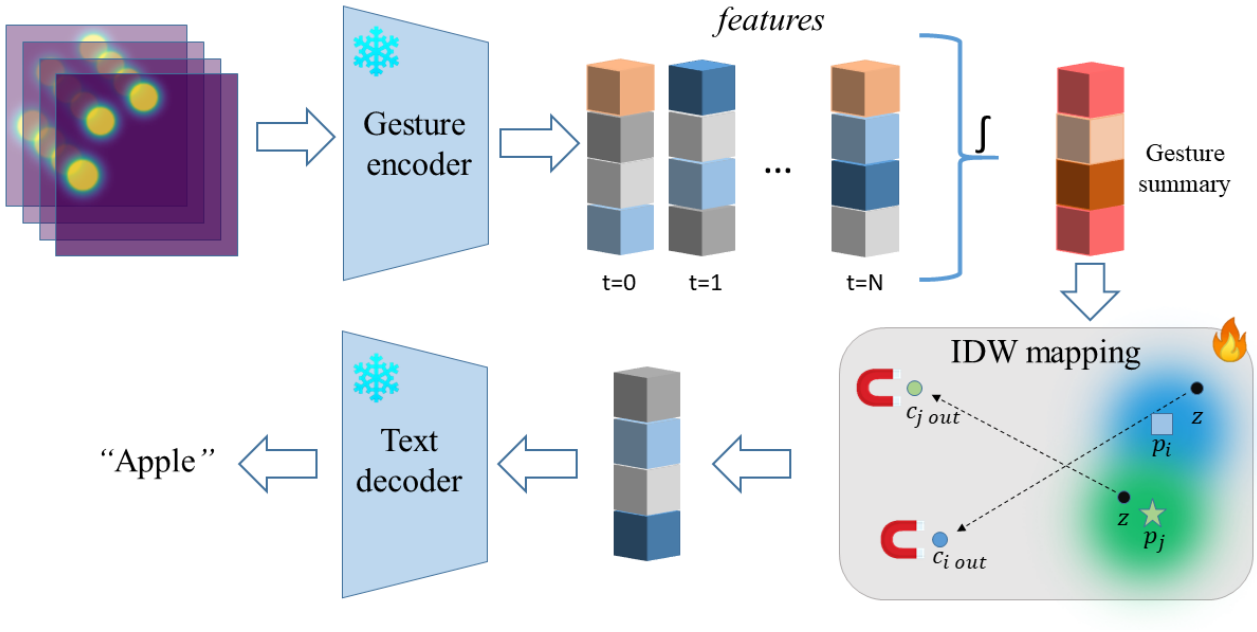

Fig. 1. End-to-end machine learning pipeline: A sequence of spatial contact frames is compressed by a convolutional Spatio-Geometric Autoencoder encoder and summarized via a path-integral temporal transformation into a fixed gesture summary vector z. An offset-stabilized non-parametric IDW mapping translates z into the linguistic latent space, where a recurrent Linguistic Sequence Autoencoder decoder reconstructs the target n-gram sequence.

The final discrete transformation is approximated as:

$$z \approx \sum_{i=1}^{N-1} \bar{p}_i \odot \cos(\omega \bar{t}_i + \phi) \Delta t_i \tag{4}$$

This path-integral approach guarantees a strict linear complexity of $O(N \cdot D)$, entirely avoiding feature explosion while remaining invariant to the sampling density of the underlying sensor data streams.

V. Non-Parametric Manifold Mapping

Once compressed into the trajectory vector z, the gesture must be cross-mapped into the target language space. We implement a non-parametric interpolation formulation designed to maintain stable predictions under input variability. While the weight distribution corresponds mathematically to Shepard's Inverse Distance Weighting (IDW) method [7], our formulation incorporates localized spatial offsets to induce a self-correcting, "magnetic" mapping effect.

The system maintains a database of localized anchor pairs (p_i, c_i) that map a known gesture summary vector p_i to its verified textual target embedding c_i. For an unknown query vector z, the system executes a k-Nearest Neighbor (k-NN) retrieval over the anchor database using a Ball Tree spatial index [8]. The interpolation weights w_j for the retrieved neighbors are computed using their inverse Euclidean distance raised to a power parameter s:

$$w_j = \frac{\frac{1}{\|z - p_j\|^s}}{\sum_{l=1}^{k} \frac{1}{\|z - p_l\|^s}} \tag{5}$$

By utilizing a high power parameter value (e.g., $s \geq 5$), the influence of distant neighbors diminishes rapidly, isolating spatial reformations to immediate local neighborhoods. The final target language vector c_{out} is generated by applying a weighted sum of translation vectors $(c_i - p_i)$ directly onto the query input z:

$$c_{\text{out}} = z + \sum_{i=1}^{k} w_i (c_i - p_i) \tag{6}$$

This offset formulation provides high structural robustness. Even if the input vector z exhibits geometric variance due to unrefined motor execution, the addition of the interpolated translation offset pulls c_{out} toward the correct target embedding c_i, preserving mapping stability under noisy conditions.

VI. Theoretical Execution and Adaptation Framework

A. Parallel Time Windows and Retroactive Segmentation

Continuous real-time gestures lack explicit demarcation boundaries. To perform segmentation dynamically, the framework outlines a theoretical multi-window system. A circular history buffer maintains recent states, from which multiple overlapping temporal windows of varying durations are generated backward from the current timestamp.

The system computes localized interpolation confidence levels across all active windows concurrently. The interface tentatively outputs the text string from the window expressing the highest instant confidence. However, if a longer temporal window subsequently reaches a higher, log-scaled confidence score—indicating a larger contextual motion provides a more

coherent linguistic interpretation—the system retroactively overrides the previous shorter prediction. Operationally, this is executed by issuing automated backspace instructions to erase the short-window characters before appending the newly validated sequence.

B. Closed-Loop Adaptation Mechanisms

To realize personalization over extended use, we propose three distinct adaptation loops within the non-parametric database:

- **Explicit Supervision**: Manual user corrections of text errors provide an explicit supervisory signal. The intended text string is encoded via the frozen Linguistic Sequence Autoencoder to extract the exact target vector c_{target}. This is paired with the corresponding trajectory summary p_{source} to instantiate or update an anchor point $(p_{\text{source}}, c_{\text{target}})$ within the Shepard database, locally correcting the space for future identical executions.
- **Implicit Reinforcement**: If a low-confidence text prediction is generated and the user continues typing without issuing corrections, the system interprets this as a passive confirmation. The algorithm reinforces the mapping by saving the pair into the database, boosting its retrieval confidence for future executions.
- **Gesture Merging**: By monitoring the operational frequency of sequential n-grams, the system can flag frequent character combinations. It permits users to combine these sequences into single, continuous compound motions. The system binds this longer continuous trajectory summary directly to the combined n-gram embedding, enabling personalized shortcuts that increase throughput.

VII. Algorithmic Validation and Synthetic Benchmarking

A. Component-Level Performance Metrics

Quantitative evaluations were performed on the individual autoencoding modules to isolate spatial and linguistic compression errors prior to evaluating the end-to-end pipeline. The Spatio-Geometric Autoencoder achieved a Pixel Accuracy of 94.46% (within a bounded ± 0.1 tolerance range) and a Mean Absolute Error (MAE) of 0.0458 across a synthetic validation set of 10,000 multi-contact topologies. This confirms clean geometric structural preservation within the 128-dimensional latent space.

The Linguistic Sequence Autoencoder demonstrated near-perfect sequence reconstruction across short-to-medium sequences, as shown in Table I. While parameter limitations within the lightweight GRU network cause degradation on extended sequences exceeding 15 characters, this trade-off was intentionally accepted to ensure immediate, deterministic inference on low-power host CPUs.

To validate the coupling of the linear temporal path integration with the offset-stabilized Inverse Distance Weighting (IDW) interpolation, a one-shot mathematical trial was executed using a standardized lexicon of 51 distinct single-stroke trajectories. The closed-loop mapping achieved a 98.03% classification accuracy (mapping 50 out of 51 test inputs onto their exact textual latent targets), verifying that the non-parametric database scales effectively without experiencing cross-talk or catastrophic forgetting. Systematic errors occurred exclusively when distinguishing highly identical geometric patterns (such as "7" vs. ">"), indicating that fine-grained separation requires higher resolution polar transformations during preprocessing.

TABLE I
Linguistic Autoencoder Sequence Reconstruction Accuracy

N-Gram Length	Char Accuracy	Exact Match	Sample Size
1–5 characters	99.99%	99.97%	3,043
6–10 characters	99.75%	99.10%	1,775
11–15 characters	96.26%	82.91%	1,703
16–20 characters	76.11%	29.07%	2,484
21+ characters	43.19%	1.71%	995

B. Online Stream Execution and Parallel Window Instability

Unlike isolated evaluation trials, real-world continuous deployment requires parsing unsegmented, real-time spatial trajectories. To evaluate this online stream environment, an operational validation protocol was constructed using a vocabulary of 32 distinct baseline gestures. These gestures utilize between 1 and 4 concurrent finger contact configurations at any given moment.

To maximize ergonomic viability, the gesture mapping was designed around intuitive, easy-to-remember motor primitives: for instance, entering the character "I" is mapped to a single-finger vertical stroke upward, while entering the character "M" is mapped to a parallel three-finger vertical stroke downward. Over extended execution cycles, the non-parametric adaptation database monitors character frequencies; it dynamically learns to merge sequentially executed gestures into single compound motions, allowing the functional vocabulary to grow organically over time.

During testing, the parallel time-window engine successfully demonstrated retroactive error correction. When a broader temporal history window reached a higher, log-scaled confidence score, the architecture successfully triggered automated backspace routines to overwrite transient, short-window predictions with the globally verified n-gram string. In practice, given sufficient execution time and user error-correction iterations, arbitrary text generation was fully achievable.

However, the continuous multi-window implementation exhibited noticeable operational instability, manifesting as stochastic character substitutions and insertion errors during rapid input. This variance is attributable to high boundary sensitivity (where minor variations in user velocity caused adjacent temporal windows to capture trailing frames from preceding movements) and density cluttering within the IDW database under high input noise.

C. Methodological Limitations

It must be explicitly acknowledged that formal empirical user testing—specifically, structured human-subject studies

across diverse user cohorts to report standardized statistical metrics like Words-Per-Minute (WPM) or total Total Error Rates (TER)—was not conducted for this iteration. The operational evaluations performed serve strictly as a mathematical and functional proof-of-concept for the underlying machine learning pipeline. Consequently, the observed typing throughput and stability are bounded by individual developer trials, and comprehensive usability profiling remains a primary objective for future implementation stages.

VIII. Conclusion and Future Work

This paper presented a layout-free machine learning framework for text entry that maps continuous, multi-finger geometric trajectories directly into a continuous linguistic space. By utilizing frozen dual latent spaces bridged by a non-parametric, offset-stabilized interpolation database, the system adapts dynamically to individual user motor habits without experiencing catastrophic forgetting. Furthermore, the linear-complexity path integral enables computationally efficient trajectory integration at a strict complexity of $O(N \cdot D)$.

While the parallel time-window segmentation architecture proved capable of running online retroactive string corrections for the 32-gesture starting vocabulary, its operational instability highlights the limitations of discrete windowing over continuous input streams. Consequently, future research will abandon discrete parallel time-window tracking in favor of a more unified temporal architecture. We intend to explore the integration of soft, attention-driven continuous temporal alignment mechanisms—such as Connectionist Temporal Classification (CTC) layers or continuous-state hidden Markov modeling—directly over the latent trajectory matrix to resolve window boundary sensitivity. Finally, we plan to execute comprehensive, multi-user empirical testing to gather formal, peer-comparable typing speed metrics and track long-term adaptation learning curves across standard human cohorts.

Acknowledgments

The author would like to thank Aleksandar Balanesko and Dimitrije Sakan for their vital contributions to the upstream hardware exploration, physical prototyping configurations, and sensor matrix data collection routines which helped inform and establish the operational constraints of this algorithmic framework.

References

[1] S. Zhai and P.-O. Kristensson, "Shorthand writing on stylus keyboard," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '03)*, New York, NY, USA: ACM, Apr. 2003, pp. 97–104.

[2] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the 38th International Conference on Machine Learning*, PMLR, Jul. 2021, pp. 8748–8763.

[3] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio, "Generating Sentences from a Continuous Space," *arXiv:1511.06349*, May 2016.

[4] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *arXiv:1312.6114*, Dec. 2013.

[5] I. Chevyrev and A. Kormilitzin, "A Primer on the Signature Method in Machine Learning," *arXiv:1603.03788*, 2016, pp. 3–64.

[6] P. Bonnier, P. Kidger, I. P. Arribas, C. Salvi, and T. Lyons, "Deep Signature Transforms," in *Advances in Neural Information Processing Systems (NeurIPS)*, May 2019.

[7] D. Shepard, "A two-dimensional interpolation function for irregularly-spaced data," in *Proceedings of the 1968 23rd ACM National Conference (ACM '68)*, New York, NY, USA: ACM, Jan. 1968, pp. 517–524.

[8] S. Omohundro, "Five Balltree Construction Algorithms," International Computer Science Institute Technical Report, 1989.