Article

Unsupervised Segmentation and Alignment of Multi-Demonstration Trajectories via Multi-Feature Saliency and Duration-Explicit HSMMs

Tianci Gao * , Konstantin A. Neusypin , Dmitry D. Dmitriev , Bo Yang , Shengren Rao

Posted Date: 22 August 2025

doi: 10.20944/preprints202508.1598.v1

Keywords: hidden semi-Markov model; learning from demonstration; unsupervised segmentation; feature fusion; topological persistence

Article

Unsupervised Segmentation and Alignment of Multi-Demonstration Trajectories via Multi-Feature Saliency and Duration-Explicit HSMMs

Tianci Gao *, Konstantin A. Neusypin, Dmitry D. Dmitriev, Bo Yang and Shengren Rao

Department IU-1 "Automatic Control Systems," Bauman Moscow State Technical University, Moscow 105005, Russian Federation

* Correspondence: gaotianci0088@gmail.com

Abstract

Learning from demonstration with multiple executions must contend with time warping, sensor noise, and alternating quasi-stationary and transition phases. We propose a label-free pipeline that couples unsupervised segmentation, duration-explicit alignment, and probabilistic encoding. A dimensionless multi-feature saliency (velocity, acceleration, curvature, direction-change rate) yields scale-robust keyframes via persistent peak–valley pairs and non-maximum suppression. A hidden semi-Markov model (HSMM) with explicit duration distributions is jointly trained across demonstrations to align trajectories on a shared semantic time base. Segment-level probabilistic motion models (GMM/GMR or ProMP, optionally combined with DMP) produce mean trajectories with calibrated covariances, directly interfacing with constrained planners. Feature weights are tuned without labels by minimizing cross-demonstration structural dispersion on the simplex via CMA-ES. Across UAV flight, autonomous driving, and robotic manipulation, the method reduces phase-boundary dispersion by 31% on UAV-Sim and by 30–36% under monotone time warps, noise, and missing data (vs. HMM); improves the sparsity–fidelity trade-off (higher time compression at comparable reconstruction error) with lower jerk; and attains nominal 2σ coverage (94–96%), indicating well-calibrated uncertainty. Ablations attribute the gains to persistence plus NMS, weight self-calibration, and duration-explicit alignment. The framework is scale-aware and computationally practical, and its uncertainty outputs feed directly into MPC/OMPL for risk-aware execution.

Keywords: hidden semi-Markov model; learning from demonstration; unsupervised segmentation; feature fusion; topological persistence

MSC: 68T05

1. Introduction

Learning from Demonstration (LfD) [1] seeks to transfer skills from a handful of expert executions to new task instances. In real deployments across aerial, driving, and manipulation domains, multiple demonstrations routinely exhibit irregular local time warping (hovering, waiting, backtracking) [2–4], sensor noise on high-dimensional signals [5,6], and alternating quasi-stationary and transition phases [7–9]. Our objective is to discover phase structure without labels, align multiple demonstrations onto a shared semantic time base, and provide calibrated, segment-wise probabilistic models that plug into constrained planners.

A central difficulty is that widely used segmentation and alignment tools either operate at the signal-threshold level or impose duration assumptions that do not reflect how humans actually perform tasks [10,11]. Threshold/template methods or simple peak detectors are brittle to sampling-rate and environment shifts; their decision surfaces move with scale changes, causing under- and over-segmentation in the presence of jitter and speed variations [12,13]. Bayesian online changepoint detection (BOCPD) relieves fixed thresholds but still hinges on hazard-rate and noise assumptions; in practice it often fragments long dwells into multiple short segments when features oscillate, and it provides no explicit duration model for downstream use [14–16]. Dynamic Time Warping (DTW) offers pairwise alignment but lacks a generative mechanism and cannot represent dwell distributions or uncertainty, which limits its utility for planning and simulation [17–19].

Latent-state models address some of these issues, yet common instantiations come with their own bias-variance trade-offs. Hidden Markov Models (HMMs) assume geometric state durations; this geometric-duration bias shortens non-geometric dwells and shifts boundaries when operators hesitate or loiter, leading to misaligned phases across trials [20]. Hidden semi-Markov Models (HSMMs) are a principled remedy because they explicitly model durations, but most prior pipelines still depend on hand-crafted feature stacks with fixed weights/hyper-parameters, trained per-demonstration or per-session [1,21,22]; robustness degrades under operator/platform shifts and tempo changes. Bayesian nonparametrics can adapt model complexity [23], but the computational overhead and stability concerns have limited their use in long-horizon, multi-demo settings.

Downstream generation has a parallel set of limitations. Dynamical Movement Primitives (DMPs) provide smooth, time-scalable execution but lack closed-form uncertainty; Gaussian Mixture Regression (GMR) and Probabilistic Movement Primitives (ProMPs) offer distributional predictions with covariances, yet both are sensitive to boundary errors—mis-segmentation inflates variances and biases means within segments [20]. More importantly, most pipelines treat segmentation/alignment and probabilistic encoding as loosely coupled stages; there is rarely a mechanism to self-calibrate features using multi-demo consistency or to learn an alignment that is simultaneously scale-aware and duration-aware [24]. Finally, while planners such as OMPL/MPC can exploit covariance for risk-aware control, few works deliver calibrated uncertainty on a shared semantic time base that is directly consumable by these planners [25].

To address these gaps, we propose a label-free pipeline that couples unsupervised segmentation, duration-explicit alignment, and probabilistic encoding into a single, self-calibrating loop. First, we compute a dimensionless multi-feature saliency by fusing velocity, acceleration, curvature, and direction-change rate; then we apply topology-aware keyframe extraction using persistent peak–valley pairs to keep only structurally significant extrema and non-maximum suppression to avoid clustered responses [26]. Second, we jointly train an HSMM across demonstrations with explicit duration distributions and an extended forward–backward recursion, producing a shared semantic time axis and phase-consistent boundaries [25]. Third, within each phase we fit probabilistic motion models—GMM/GMR or ProMP, optionally combined with DMP for execution—to obtain mean trajectories with calibrated covariances [20]. Crucially, we close the loop by learning the saliency weights without labels: a CMA-ES [27] search on the probability simplex minimizes cross-demonstration structural dispersion, automatically rebalancing features so that segmentation and alignment are mutually consistent. Compared with BOCPD, thresholding, or HMM-based pipelines, this design is explicitly duration-aware, scale/time-warp robust, and planner-ready with uncertainty that is calibrated and comparable across operators.

Contributions.

1. **Scale-/time-warp-robust saliency.** A topology-aware, multi-feature saliency (persistence and non-maximum suppression) that stabilizes keyframes under noise and tempo variations, yielding **sparser yet more stable** anchors than signal-level detectors.
2. **Joint HSMM alignment with explicit durations.** Cross-demo training with extended forward–backward/Viterbi recursions; model order selected by a joint criterion combining BIC and alignment error (AE) to balance fit and parsimony.
3. **Label-free feature-weight self-calibration.** CMA-ES on the weight simplex to **minimize cross-demo structural dispersion**, eliminating hand-tuned fusion and improving phase consistency.

4. **Calibrated probabilistic encoding for planning.** Segment-wise GMR/ProMP (optionally fused with DMP) returning **means and covariances** that integrate directly with **OMPL/MPC** for risk-aware execution.

Empirically, on UAV flight, autonomous-driving, and manipulation benchmarks, our method reduces phase-boundary dispersion by ≈31% on UAV-Sim and by 30–36% under monotone time warps, additive noise, and missing data compared with HMM variants; it improves the sparsity–fidelity–smoothness trade-off (higher time compression at comparable reconstruction error with lower jerk) and achieves nominal 2σ coverage (94–96%), indicating well-calibrated uncertainty. Section 2 details the pipeline; Section 3 reports datasets, metrics, baselines (including BOCPD and HMM), ablations, and robustness; Section 4 discusses positioning, limitations, and future work; Section 5 concludes.

2. Materials and Methods

Figure 1 sketches the end-to-end workflow that maps multiple demonstrations to executable trajectories under physical and safety constraints: (i) **unsupervised segmentation** from multi-feature saliency with topology-aware keyframes; (ii) **duration-explicit alignment** across demonstrations via a hidden semi-Markov model (HSMM) with shared parameters; and (iii) **probabilistic encoding** of each phase (GMM/GMR or ProMP, optionally combined with DMP) producing mean trajectories with calibrated covariances. Outputs include **phase-consistent labels**, **segment-wise probabilistic models**, and **constraint-aware executable trajectories**.

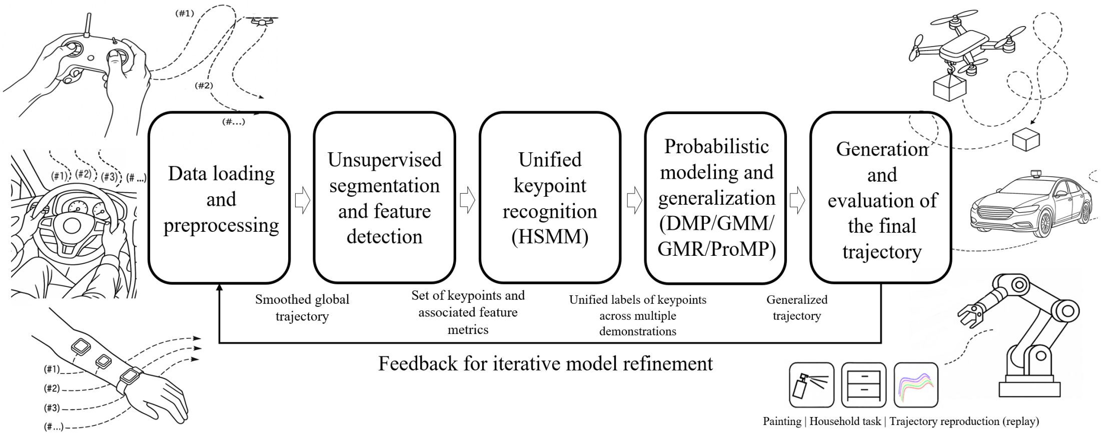

(#1)
(#2)
(#...)
(#1) (#2) (#3) (# ...)
(#1)
(#2)
(#3)
(#...)
Data loading and preprocessing
Unsupervised segmentation and feature detection
Unified keypoint recognition (HSMM)
Probabilistic modeling and generalization (DMP/GMM/GMR/ProMP)
Generation and evaluation of the final trajectory
Smoothed global trajectory
Set of keypoints and associated feature metrics
Unified labels of keypoints across multiple demonstrations
Generalized trajectory
Feedback for iterative model refinement
Painting | Household task | Trajectory reproduction (replay)

Figure 1. Overview of the proposed pipeline from multiple demonstrations to executable trajectories.

2.1. Inputs, Outputs, and Assumptions

Inputs.

We observe M independent demonstrations (optionally multi-modal), $\{P^{(m)}\}_{m=1}^{M}, P^{(m)} = \{P^{(m)}(t)\}_{t=1}^{T_m}$, sampled with fixed period Δt. When available, auxiliary channels (e.g., pose, force/tactile, depth) are concatenated into the observation vector used downstream.

Outputs.

1. A shared set of semantic phases $\{S_k\}_{k=1}^{N^\star}$ and per-demo boundaries $\{\tau_k^{(m)}\}$;
2. For each phase, a segment-wise generative model—DMP, GMM/GMR, or ProMP, alone or in combination—returning mean trajectories and covariance estimates;
3. Under constraints $\mathcal{C}$ (terminal/waypoints, velocity/acceleration limits, etc.), an executable trajectory and associated risk measures computed from covariances.

Assumptions.

1. Demonstrations consist of alternating quasi-stationary and transition segments.
2. Time deformation is order-preserving (the semantic phase order does not change).

3. Observation noise is moderate and can be mitigated by local smoothing and statistical filtering.

2.2. Multi-Feature Analysis and Automatic Segmentation

2.2.1. Data Ingestion and Pre-modulation

Let a single demonstration be the discrete sequence $\mathcal{D} = \{\Delta\mathbf{x}_t\}_{t=1}^{T}, \Delta\mathbf{x}_t \in \mathbb{R}^d$, where d is the number of observed degrees of freedom (e.g., $d = 3$ for UAV or mobile bases; $d = 6$ for industrial manipulators). Cumulative sums yield the Cartesian trajectory of the tool center point

$$\mathbf{P}(t) = \left[p_x(t), p_y(t), p_z(t)\right]^{\top}, t = 1, \dots, T, \tag{2.1}$$

We apply a one-dimensional **Savitzky–Golay** local polynomial smoother to $\mathbf{P}(t)$ before computing derivatives, suppressing high-frequency tele-operation jitter and stabilizing numerical differentiation. The filter is equivalent to an $(r + 1)$-order Taylor truncation in time and ensures

$$\|\widetilde{\mathbf{P}} - \mathbf{P}\|_{\infty} \leq c_W \varepsilon, c_W < 1,$$

where ε bounds the measurement noise and the window $W \in \{5,7,9\}$. Unless stated otherwise, $\widetilde{\mathbf{P}}$ denotes the smoothed trajectory [28].

Implementation note. Central differences are used for derivatives; all channels are time-synchronized at a fixed sampling period Δt.

2.2.2. Feature Computation and Saliency Fusion

On the smoothed trajectory we compute four complementary, time-varying features that reveal when, where, and how kinematic changes occur ($\Delta t \left(= \frac{1}{f_s}\right)$ is fixed).

1. **Velocity.** Let

$$\mathbf{v}(t) = \frac{\mathbf{P}(t+1) - \mathbf{P}(t)}{\Delta t}, v(t) = \|\mathbf{v}(t)\|_2. \tag{2.2}$$

2. **Acceleration.**

$$\mathbf{a}(t) = \frac{\mathbf{v}(t+1) - \mathbf{v}(t)}{\Delta t}, a(t) = \|\mathbf{a}(t)\|_2. \tag{2.3}$$

Peaks indicate abrupt speed changes.

3. **Curvature.** With ΔP(t)=P(t+1)-P(t),

$$\kappa(t) = \frac{2\|\Delta\mathbf{P}(t-1) \times \Delta\mathbf{P}(t)\|}{\|\Delta\mathbf{P}(t-1)\|\|\Delta\mathbf{P}(t)\|\|\Delta\mathbf{P}(t) - \Delta\mathbf{P}(t-1)\|} \tag{2.4}$$

Curvature measures spatial bending and is naturally invariant to global time scaling due to the cubic velocity term [29,30].

4. **Direction-Change Rate (DCR).** Define the unit-direction vector $\hat{\mathbf{v}}(t) = \frac{\mathbf{v}(t)}{\|\mathbf{v}(t)\|_2}$. To avoid numerical issues at very low speed, introduce a threshold $v_{min} > 0$ and set

$$\mathrm{DCR}(t) = \begin{cases} \|\hat{v}(t) - \hat{v}(t-1)\|_2, & \|v(t)\|_2 \geq v_{\min} \\ 0, & \|v(t)\|_2 < v_{\min} \end{cases} \tag{2.5}$$

5. **Dimensionless fusion.** Apply min–max normalization to each feature to obtain $\tilde{v}(t), \tilde{a}(t), \tilde{\kappa}(t), \overline{\mathrm{DCR}}(t) \in [0,1]$. For a weight vector $w = [w_v, w_a, w_\kappa, w_d]^{\top}$ with $w_i \geq 0$ and $\sum_i w_i = 1$, define the fused saliency

$$\mathrm{Score}(t; w) = w_v \tilde{v}(t) + w_a \tilde{a}(t) + w_\kappa \tilde{\kappa}(t) + w_d \widetilde{\mathrm{DCR}}(t) \tag{2.6}$$

The weights w are learned **without labels** in Section 2.2.4.

2.2.3. Keyframe Extraction with Topological Simplification

The saliency $\mathrm{Score}(t)$ compresses multi-source information into a 1-D signal, but segmentation should rely on structural extrema (global landmarks), not every minor fluctuation. We adopt a bottom-up screening that contracts a dense set of local extrema into a sparse, stable keyframe set [1,31].

1. **Candidate extrema via quantile thresholds.**

Let $Q = \{q_1, \dots, q_L\} \subset (0,1)$be a grid of quantiles (e.g., uniformly in [0.60,0.95]). For each $q \in Q$:

- set $\tau_q = \ \mathit{quantile}_q(\ \texttt{Score}\ )$;
- collect indices $\widetilde{\mathrm{E}}_q = \{t \mid \mathrm{Score}(t) > \tau_q\}$ and snap each t to the nearest local extremum within a radius-3 neighborhood.

To pick a unique $q^\star$, minimize a **sparsity–fidelity** loss

$$L(q) = \underbrace{\left|\tilde{E}_q\right|/T}_{\texttt{sparsity}} + \lambda \underbrace{\mathrm{MSE}\left(P, \hat{P}_{\tilde{E}_q}\right)}_{\texttt{reconstruction}},$$

where $\hat{P}_{\widetilde{\mathrm{E}}_q}$ is the spline reconstruction at $\widetilde{\mathrm{E}}_q$ and $\lambda > 0$ reflects the admissible reconstruction error. Set $q^\star = \arg \underset{q \in Q}{min} \mathcal{L}(q), \widetilde{\mathrm{E}} = \widetilde{\mathrm{E}}_{q^\star}$.

2. **Persistence thresholding (scale-invariant importance).**

For adjacent peak–valley pairs (p_{max}, p_{min}) of $\mathrm{Score}(t)$, define the **persistence**

$$\mathrm{pers}(p_{max}, p_{min}) = |\mathrm{Score}(p_{max}) - \mathrm{Score}(p_{min})| \qquad (2.7)$$

Small persistence typically indicates noise or micro-tremor; large persistence corresponds to genuine kinematic transitions. Plot

$$g(\alpha) = \mid \{(p_{max}, p_{min}) \mid \ \mathrm{pers}\ > \alpha\} \mid,$$

which empirically exhibits a plateau–cliff–stable pattern; the elbow $\alpha^\star$ is detected by Kneedle. Keep

$$\mathrm{E}^\dagger = \{p_i \mid \mathrm{pers}(p_i) >\}$$

Because persistence depends only on amplitude differences, it is invariant to vertical scaling and mild time stretching, facilitating cross-demo comparability [32,33].

3. **Non-maximum suppression (NMS).**

To avoid peak clustering, scan $\mathrm{E}^\dagger$ with a sliding window of NMS_w frames and retain an extremum only if it is the largest (same polarity) within the window. The final keyframe set is

$$\mathcal{K} = \mathrm{NMS}_w(\mathrm{E}^\dagger)$$

The effect of persistence thresholding and NMS is illustrated in Figure 2, which contracts dense peak clusters into a sparse, stable keyframe set $\mathcal{K}$. The black solid curve is the original saliency signal $\mathrm{Score}(t)$; the dashed curve is the simplified signal after persistence-based filtering (elbow $\alpha^\star$ detected by Kneedle) and non-maximum suppression (window $\boldsymbol{w}$). Green crosses mark the final retained keyframes $\mathcal{K}$. The procedure removes clustered minor extrema while preserving structural landmarks that drive segmentation and alignment.

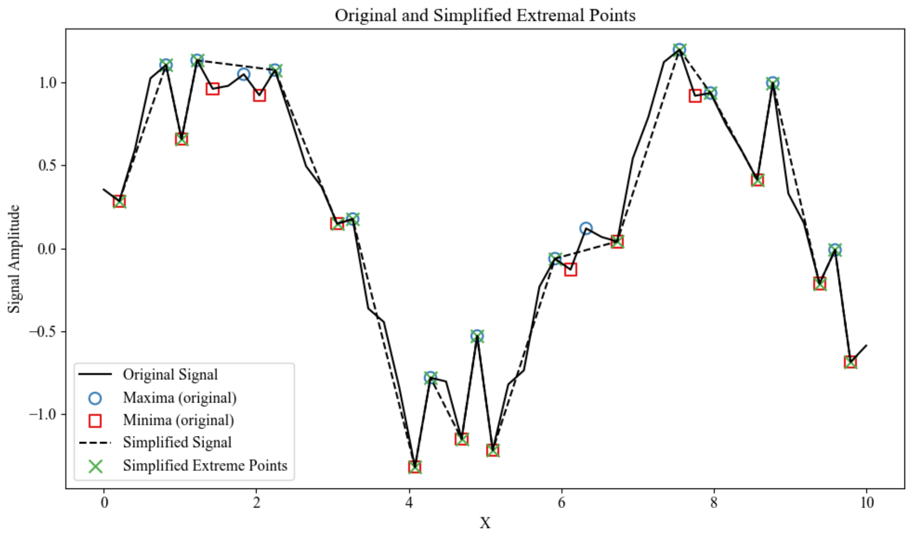


Figure 2. Topological simplification and keyframe selection (persistence + NMS).

With $\mathcal{K}$ in place, the reliability of subsequent segmentation and HSMM alignment still depends on the saliency weights w; Section 2.2.4 details a label-free, consistency-driven calibration.

2.2.4. Adaptive Feature-Weight Learning

In (2.6), the saliency $\text{Score}(t; w)$ is a linear fusion of four heterogeneous features. Assigning heuristic, fixed weights $w = [w_v, w_a, w_k, w_d]^\top$ typically overfits a particular operator, platform, or task, degrading segmentation under distribution shift. We therefore treat w as model parameters to be estimated without labels by enforcing cross-demonstration consistency.

1. **Consistency functional.**

For M demonstrations $\{\mathbf{P}^{(m)}\}_{m=1}^{M}$ and a candidate w, define

$$\mathcal{F}(\cdot; \boldsymbol{w}): \mathbf{P}^{(m)} \longmapsto \widehat{\mathbf{P}}^{(m)}(\boldsymbol{w}), \tag{2.8}$$

denote the composition of saliency construction, keyframe extraction (Figure 2), HSMM-based alignment (Section 2.3), and resampling on shared semantic nodes. If w is well chosen, the reconstructions $\{\hat{P}^{(m)}(w)\}$ should be congruent in shape and timing. We quantify this by the mean point-to-point structural dispersion.

$$\text{SOD}(\boldsymbol{w}) = \frac{2}{M(M-1)} \sum_{1 \le m < m' \le M} \frac{1}{T_{min}} \sum_{t=1}^{T_{min}} \left\| \hat{\mathbf{p}}_t^{(m)}(w) - \hat{\mathbf{p}}_t^{(m')}(w) \right\|_2, \tag{2.9}$$

here T_{min} is the minimal resampled length. Smaller SOD means lower structural variance on the shared semantic time base.

2. **Objective and constraints.**

To prevent dominance by any single channel and to improve identifiability, we regularize with a light ℓ_2 penalty and minimize over the probability simplex:

$$w^\star = \arg \min_{\boldsymbol{w} \in \Delta^3} (\text{SOD}(\boldsymbol{w}) + \lambda \|\boldsymbol{w}\|_2^2), \Delta^3 = \left\{ \boldsymbol{w} \mid w_i \ge 0, \sum_i w_i = 1 \right\}, \tag{2.10}$$

Because F contains non-smooth steps (extrema detection, discrete decoding), $\mathbf{SOD}(\boldsymbol{w})$ is non-differentiable and gradient methods are unreliable.

3. **Solver and feasibility.**

We employ the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) in an unconstrained space together with a softmax reparameterization

$$w_i(z) = \frac{\exp(z_i)}{\sum_{j=1}^{4} \exp(z_j)}, z \in \mathbb{R}^4$$

which enforces $w \in \Delta^3$ at every iteration while preserving Gaussian updates in z-coordinates. CMA-ES is appropriate here because it (i) requires only function values of $J(w) = \text{SOD}(w) + \lambda \|w\|_2^2$; (ii) adapts search scales via the evolving covariance; and (iii) is empirically robust on multi-modal, non-convex objectives [34,35]. We terminate when either the covariance trace shrinks below a preset ratio (indicating a local neighborhood) or the successive improvement $|J^{(k)} - J^{(k-1)}|$ falls below ε. Under mild regularity on J in the z-space, the iterates approach first-order stationarity (a near-KKT solution) with high probability [36].

4. **Computational profile.**

Each evaluation of $J(w)$ entails one pass of keyframe extraction (linear in sequence length T) and one HSMM forward–backward/decoding pass per demonstration with complexity $O(NTD_{max})$; overall cost scales linearly in M. For numerical stability we implement recursions in the log domain (log-sum-exp), apply a diagonal floor to GMM covariances, and cache feature streams across outer-loop calls.

2.3. *Multi-Demo Alignment and Segmentation via a Duration-Explicit HSMM*

Given the sparse, scale-invariant keyframes $\mathcal{K}^{(m)} = \left\{t_k^{(m)}\right\}_{k=1}^{K_m}$ obtained from saliency (Section 2.2) and the learned fusion weights (Section 2.2.4), we seek a **shared semantic time base** across demonstrations. Directly matching wall-clock indices is unreliable due to operator-dependent pauses and backtracking; instead, we align demonstrations probabilistically via a **Hidden semi-Markov model (HSMM)** that explicitly models state durations (Figure 3) [37,38].

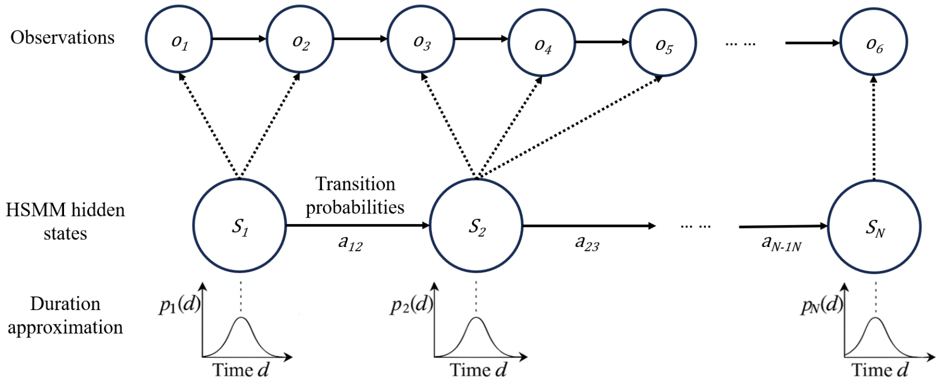


Figure 3. Schematic of the HSMM. Circles denote random variables (top: observations o_t; bottom: latent phases S_i). Solid arrows indicate Markovian dependencies; dashed links connect latent states to observation likelihoods. Shaded curves at the bottom depict state-specific duration distributions $p_i(d)$.

2.3.1. Model and Generative Mechanism

Let $\mathcal{S} = \{S_1, S_2, \dots, S_N\}, N \ll K_m$ be latent phases, each representing a macro-action (e.g., grasp, insert, lift). With parameters $\Theta = \{\pi, A, \{p_i\}, \{b_i\}\}$, the HSMM generates $o_t \in \mathbb{R}^d$ at $t = 1, \dots, T$ as follows:

a) **Initial phase:** $q_1 \sim \mathrm{Cat}(\pi), \sum_i \pi_i = 1$.

b) **Duration:** for the current phase q_k, sample dwell length $d_k \sim p_{q_k}(d) \in \mathbb{N}_{>0}$.

c) **Observations:** for $\tau_{k-1} < t \le \tau_k$ with $\tau_k = \sum_{j \le k} d_j$, $o_t \sim b_{q_k}(\cdot), b_i(o) = \sum_{m=1}^{M_i} \pi_{i,m} \mathcal{N}\left(o \mid \mu_{i,m}, \Sigma_{i,m}\right)$

d) **Transition:** $\Pr(q_{k+1} = j \mid q_k) = a_{q_k} j$; terminal transitions end the sequence.

Writing $O = (o_1, \dots, o_T), Q = (q_1, \dots, q_n), D = (d_1, \dots, d_n)$, and $q^*(t) = q_k$ for $\tau_{k-1} < t \le \tau_k$, the joint density is

$$P(O, Q, D) = \pi_{q_1} p_{q_1}(d_1) \prod_{k=1}^{n-1} \left[a_{q_k q_{k+1}} p_{q_{k+1}}(d_{k+1})\right] \prod_{t=1}^{T} b_{q^*(t)}(o_t). \tag{2.11}$$

Observation design. We concatenate kinematic descriptors (e.g., curvature, speed magnitude, DCR) and any available modalities (pose, force/tactile, depth) into o_t.

Duration choices. We use either discrete $p_i(d)$ with support 1: D_{max} , or truncated Gaussian/Gamma families to accommodate unequal dwell times [39].

2.3.2. Parameter Estimation: Extended Baum–Welch

We maximize the total log-likelihood over demonstrations

$$\mathcal{L}(\Theta) = \sum_{m=1}^{M} \log P\left(O^{(m)} \mid \Theta\right). \tag{2.12}$$

Explicit durations break first-order Markovity; forward–backward recursions therefore enumerate duration indices.

1. **Forward variable (leaving i at time t).**

$$\alpha_t(i)=\sum_{j=1}^{N}\sum_{d=1}^{min(D_{max},t)}\alpha_{t-d}(j)a_{ji}p_i(d)\prod_{r=0}^{d-1}b_i(o_{t-r}),\alpha_0(i)=\pi_i. \tag{2.13}$$

2. **Backward variable.**

$$\beta_t(i)=\sum_{j=1}^{N}\sum_{d=1}^{T-t}\left[a_{ij}p_j(d)\prod_{r=0}^{d-1}b_j(o_{t+1+r})\beta_{t+d}(j)\right],\beta_T(i)=1. \tag{2.14}$$

3. **Posteriors (E-step).**

$$\gamma_t(i)=\frac{\sum_{d=1}^{min(D_{max},t)}\sum_{r=0}^{d-1}\alpha_{t-r}(i)p_i(d)\prod_{u=0}^{d-1}b_i(o_{t-r+u})\beta_{t-r}(i)}{\sum_{k=1}^{N}\alpha_T(k)} \tag{2.15}$$

$$\xi_{t,d}(i,j)=\frac{\alpha_{t-d}(i)a_{ij}p_j(d)\prod_{r=0}^{d-1}b_j(o_{t-r})\beta_t(j)}{\sum_{k=1}^{N}\alpha_T(k)} \tag{2.16}$$

Sum γ,ξ over m to obtain corpus-level sufficient statistics.

4. **M-step (closed forms).**

$$\begin{aligned}\pi_i^{\text{new}}&=\gamma_{t=1}(i),\\ a_{ij}^{\text{new}}&=\frac{\sum_{t,d}\xi_{t,d}(i,j)}{\sum_t\gamma_t(i)},\\ p_i^{\text{new}}(d)&=\frac{\sum_t\xi_{t,d}(i,\cdot)}{\sum_{t,d'}\xi_{t,d'}(i,\cdot)},\\ \mu_{i,m}^{\text{new}}&=\frac{\sum_t\gamma_{t,m}(i)o_t}{\sum_t\gamma_{t,m}(i)},\\ \Sigma_{i,m}^{\text{new}}&=\frac{\sum_t\gamma_{t,m}(i)(o_t-\mu_{i,m})(o_t-\mu_{i,m})^{\top}}{\sum_t\gamma_{t,m}(i)}\end{aligned} \tag{2.17}$$

with $\gamma_{t,m}(i)=\gamma_t(i)\pi_{i,m}\mathcal{N}(o_t\mid\mu_{i,m},\Sigma_{i,m})/b_i(o_t)$

5. **Numerical stability.**
All recursions are implemented in the **log domain** using **log-sum-exp**

$$\mathrm{LSE}(x_1,\ldots,x_K)=\log\sum_{k=1}^{K}e^{x_k}, \tag{2.18}$$

and GMM covariances receive a diagonal floor δI. EM monotonicity guarantees convergence of$\mathcal{L}^{(k)}$ to a stationary point [40,41].

2.3.3. Semantic Time Axis: Decoding and Outputs

After convergence, **Viterbi decoding** yields the MAP phase path $\hat{Q}^{(m)}=\left(q_1^{(m)},\ldots,q_N^{(m)}\right)$ and durations $\widehat{D}^{(m)}=\left(d_1^{(m)},\ldots,d_N^{(m)}\right)$ for each demonstration, with $\sum_{k=1}^{N}d_k^{(m)}=T_m$. Define cumulative boundaries $\tau_k^{(m)}=\sum_{j=1}^{k}d_j^{(m)},k=1,\ldots,N$, so segments $\left(\tau_{k-1}^{(m)},\tau_k^{(m)}\right]$ correspond to the same semantic phase S_k **across** demonstrations. Figures 4 and 5 illustrate 3-D keyframe locations and HSMM reconstructions aligned to salient kinematic transitions.

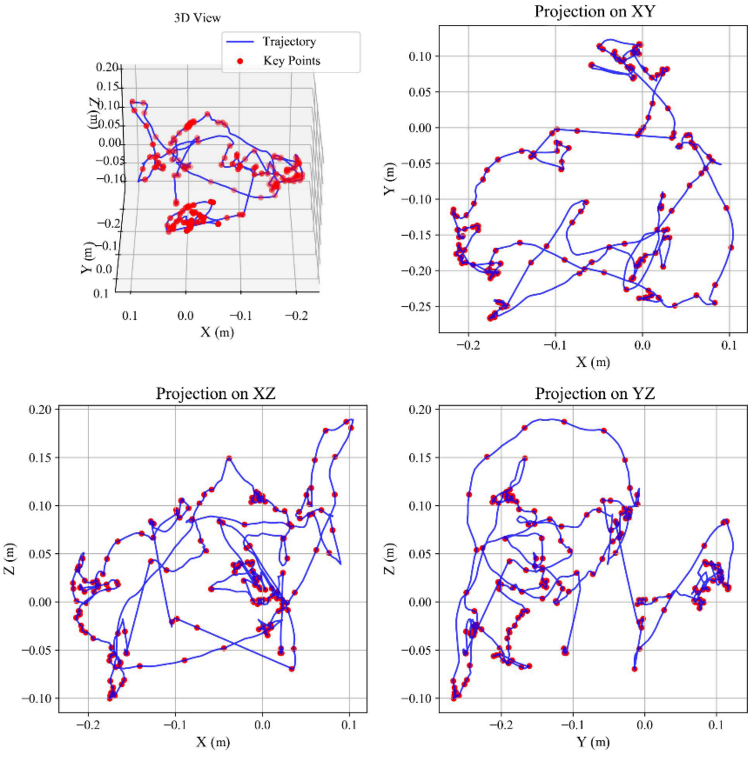


Figure 4. Example 3D trajectories with keyframes (red dots) shown in three orthographic projections. Keyframes concentrate at motion-structure bends.

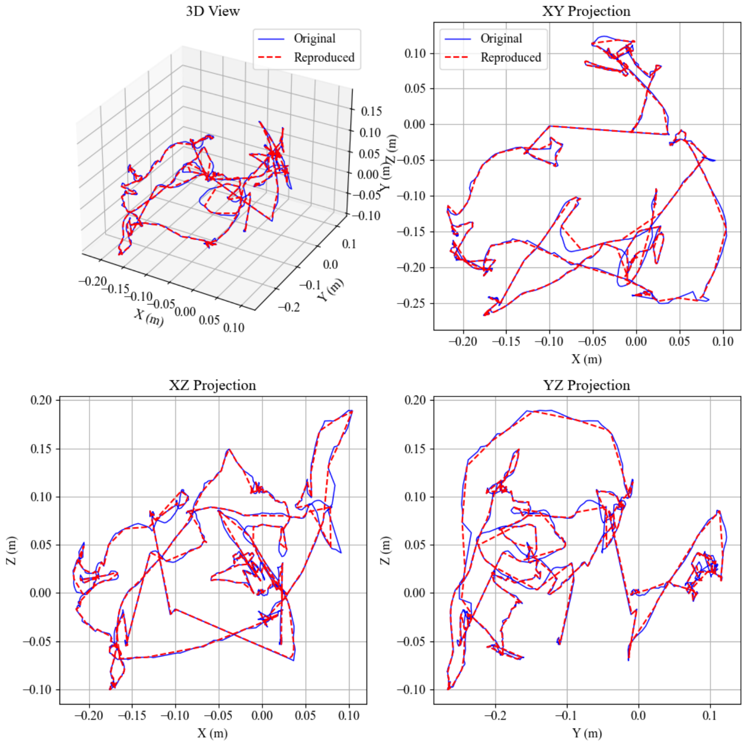


Figure 5. Original trajectory (blue) versus HSMM reconstruction (red dashed). Boundaries align with salient kinematic transitions, validating the learned semantic phases.

2.3.4. Alignment Quality, Model Selection, and Robustness

Alignment metric. We quantify cross-demo temporal agreement by

$$\mathrm{AE}(N)=\sum_{m=1}^{M}\left\|\left(\tau_1^{(m)},\dots,\tau_N^{(m)}\right)-\left(\bar{\tau}_1,\dots,\bar{\tau}_N\right)\right\|_2,\bar{\tau}_k=\frac{1}{M}\sum_{m}\tau_k^{(m)}. \tag{2.19}$$

Model selection. To balance fit and parsimony, we use

$$\mathrm{BIC}(N)=-2\log\,\mathcal{L}(\Theta_N^{\star})+\kappa_N\log\left(\sum_{m}T_m\right), \tag{2.20}$$

and select $N^{\star}$ by jointly considering $\mathrm{AE}(N)$ and $\mathrm{BIC}(N)$.

Robustness to time warps. For any order-preserving reparameterization of time, $t\mapsto\tau(t)$, the decoded phase *order* $(S_1,\dots,S_N)$ is invariant; only the tail behavior of duration distributions $p_i(d)$ is rescaled. This accounts for operator-dependent slowdowns, hesitations, or hovering.

Complexity. Keyframe processing is $O(T)$ per sequence; duration-explicit forward–backward is $O(NTD_{max})$; training scales linearly in the number of demonstrations M. Outputs include $\{S_k\}$, and segment-wise statistics for direct use by planners (e.g., OMPL/MPC).

2.4. Statistical Motion Primitives and Probabilistic Generation

With cross-demonstration phases and boundaries $\{\mathcal{S}_k\}_{k=1}^{N}$, $\left\{\tau_k^{(m)}\right\}$ obtained from the duration-explicit HSMM (Section 2.3), we model each phase *within* the shared semantic time base. For demonstration m, let

$$\mathcal{C}_k^{(m)}=\left\{P^{(m)}(t);t\in\left(\tau_{k-1}^{(m)},\tau_k^{(m)}\right]\right\}$$

denote the spatio-temporal segment for phase $\mathcal{S}_k$. Our objective is a segment-wise generative model that (i) captures cross-demo variability, (ii) supports conditioning and duration re-scaling, and (iii) yields calibrated uncertainty for downstream planning and safety assessment. We instantiate three complementary families: DMP, GMM/GMR, and ProMP. Unless otherwise noted, time within a segment is normalized to a phase variable $s\in[0,1]$, ensuring a uniform interface across models.

2.4.1. Dynamic Movement Primitives (DMP)

1. **Single-segment dynamics.**

For a one-degree-of-freedom trajectory $x(t)$, the classical DMP represents motion as a critically damped second-order system with a nonlinear forcing term:

$$\begin{aligned}\tau\dot{z}&=\alpha_z(\beta_z(g-x)-z)+f(s),\\ \tau\dot{x}&=z,\tau\dot{s}=-\alpha_s s,\end{aligned} \tag{2.21}$$

where g is the segment goal, $s\in(0,1]$ is a phase variable, and

$$f(s)=\frac{\sum_{i=1}^{B}w_i\psi_i(s)}{\sum_{i=1}^{B}\psi_i(s)}(g-x_0),\psi_i(s)=\exp\left[-h_i(s-c_i)^2\right] \tag{2.22}$$

Given $\tau,\alpha_z,\beta_z,\alpha_s$, the weights $\{w_i\}$ are obtained by least squares (or locally weighted regression). Multi-DOF trajectories are modeled component-wise or via task-space coupling.

2. **Segment coupling and smoothness.**

Let $\bar{d}_k$ be the mean duration of phase S_k across demonstrations; set $\tau_k=\bar{d}_k\Delta t$. We fit $\left\{w_i^{(k)}\right\}_{i=1}^{B}$ per segment and compose $\{(\mathrm{DMP}_k,\tau_k)\}_{k=1}^{N}$ along the decoded boundaries. Because the state(x,z) is continuous across boundaries, the concatenation is C^1 **-continuous** without auxiliary velocity/acceleration matchers. DMPs thus offer low jerk and robust time-scaling at execution.

2.4.2. Gaussian Mixture Modeling and Regression (GMM/GMR)

Mixture modeling. For each segment, we pair normalized phase with position, $(s_t,y_t)\in\mathbb{R}^{1+d}$, and fit a K-component mixture

$$p(\mathbf{s},\mathbf{y}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \Sigma_k) \qquad (2.23)$$

typically with diagonal (or block-diagonal) Σ_k to denoise while retaining principal correlations.

Regression and uncertainty. At execution, Gaussian Mixture Regression (GMR) produces the conditional

$$p(\mathbf{y} \mid \mathbf{s}) = \mathcal{N}(\hat{\mu}(\mathbf{s}), \hat{\Sigma}(\mathbf{s})),$$

yielding a closed-form mean trajectory and a **phase-indexed covariance band**. Analytical derivatives of $\hat{\mu}(\mathbf{s})$ and $\hat{\Sigma}(\mathbf{s})$ facilitate integration with MPC for online terminal corrections and constraint handling.

2.4.3. Probabilistic Movement Primitives (ProMP)

Bayesian representation. Using basis functions $\Phi_t = [\phi_1(t), \dots, \phi_B(t)]^\top$, a segment trajectory is modeled as

$$\boldsymbol{y}_t = \Phi_t^\top \boldsymbol{w} + \varepsilon, \varepsilon \sim \mathcal{N}\big(\boldsymbol{0}, \Sigma_y\big), \boldsymbol{w} \sim \mathcal{N}(\mu_w, \Sigma_w) \qquad (2.24)$$

where (μ_w, Σ_w) encode the distribution over shapes learned from multiple demonstrations (EM or closed-form under Gaussian assumptions).

Conditioning and coordination. Linear constraints—endpoints, waypoints, or partial observations—are imposed by conditioning the weight distribution:

$$w \mid (CY_T = g) \sim \mathcal{N}(\mu_w + K_T(g - \Phi_T \mu_w), \Sigma_w - K_T \Phi_T \Sigma_w), \qquad (2.25)$$

with $K_T = \Sigma_w \Phi_T (\Phi_T^\top \Sigma_w \Phi_T + \Sigma_\varepsilon)^{-1}$. Sampling from this posterior and stitching segments at decoded boundaries yields constraint-consistent trajectories with **explicit predictive uncertainty**.

2.4.4. Model Choice and Complementarity

No single primitive dominates: DMP excels in real-time low-jerk execution; GMM/GMR gives closed-form means/covariances and gradients; ProMP supports exact linear-Gaussian conditioning for multi-goal tasks. Table 1 summarizes properties. Figure 6 shows HSMM-aligned ProMP generation with calibrated uncertainty bands:

- **DMP** excels at real-time execution and low jerk with simple time scaling;
- **GMM/GMR** offers closed-form means and covariances over phase and is convenient for planners needing analytic gradients;
- **ProMP** provides a distribution over shapes with exact linear-Gaussian conditioning, ideal for multi-goal tasks and collaboration.

Table 1. Segment-Level Motion Primitives: Properties and Trade-offs.

Property	DMP	GMM/GMR	ProMP
Shape representation	Basis functions + 2nd-order stable system	Global Gaussian mixture over (s, y)	Gaussian over weights w
Duration adaptation	Via τ_k time scaling	Requires resampling in phase	Basis-phase re-timing
Uncertainty	No closed-form (MC if needed)	Analytic $\hat{\Sigma}(s)$	Analytic posterior over w
Online constraints	Endpoints/velocities easy	Refit or constrained regression	Exact linear-Gaussian conditioning
Execution smoothness	Low jerk (native dynamics)	Depends on mixture fit	Depends on basis and priors

In practice, we first fit **GMM/GMR** to obtain mean/covariance; use the GMR mean to **initialize DMP** weights for smooth execution; and, when hard/soft constraints or multi-goal adaptation are required, **overlay ProMP conditioning** on top of the DMP nominal to reconcile smoothness with constraints.

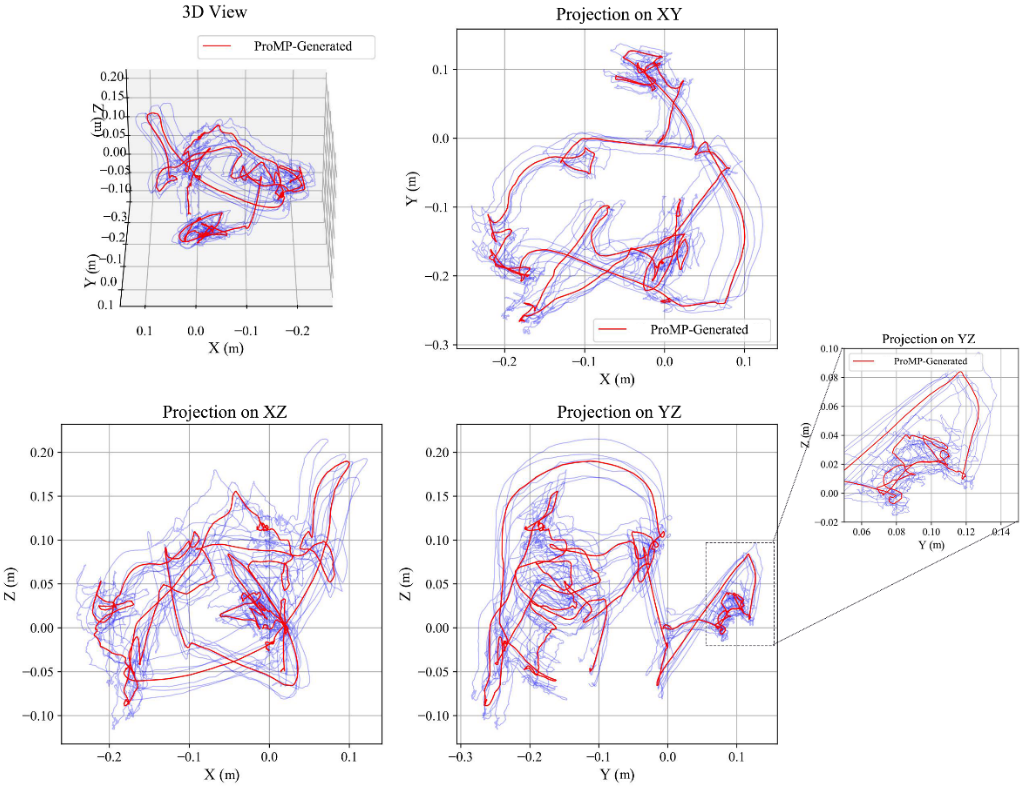


Figure 6. HSMM-aligned ProMP generation. Blue curves show multiple aligned demonstrations; red curves depict ProMP samples under a new terminal constraint; translucent bands indicate the 2σ credible region. The two-level probabilistic structure (HSMM over phases; ProMP within segment) achieves global timing consistency with locally adjustable motion.

Interface to planning and safety. Segment-wise covariances (from GMR or ProMP) propagate to risk metrics and constraint tightening in MPC/OMPL. Because segments share a **semantic** time base, the uncertainty is comparable across operators and platforms, enabling principled safety margins (e.g., 2σ envelopes) and priority-aware blending when composing multi-segment tasks.

Computational profile. Per segment, DMP fitting is linear in the number of basis functions; GMM/GMR training is $O(KT)$ per EM iteration; ProMP estimation is closed-form/EM with state dimension B. Since segments are independent given HSMM alignment, training **parallelizes over phases** and scales linearly with the number of demonstrations.

3. Experiments and Results

3.1. Objectives and Evaluation Protocol

This section subjects the proposed end-to-end pipeline to a **cross-domain, reproducible** evaluation covering the three core components introduced in Section 2: (i) **unsupervised segmentation** (multi-feature saliency with topological persistence), (ii) **duration-explicit alignment** (HSMM), and (iii) **probabilistic in-segment generation** (GMM/GMR, ProMP, optionally DMP). We target three complementary questions:

1. **Segmentation robustness.** Do multi-feature saliency and topological persistence yield **sparse yet structurally stable** keyframes under heterogeneous noise and tempo variations?

2. **Semantic alignment quality.** Does **duration-explicit** HSMM reduce cross-demonstration time dispersion when non-geometric dwelling is present (e.g., hover, wait)?
3. **Generator calibration.** On the shared semantic time base, do segment-wise probabilistic models achieve **low reconstruction error**, **nominal uncertainty coverage**, and **dynamically schedulable** executions?

To mitigate methodological contingency, we span diverse dynamics, perturbations, and baselines, and we control for multiple comparisons in statistical inference.

3.2. Tasks, Datasets, and Testable Hypotheses

We consider three representative domains:

- **Domain A—UAV-Sim (multi-scene flight).** 100 Hz sampling. Subtasks include *take-off–lift–cruise–drop* and *gate-pass–loiter–gate-pass*. Six subjects, 20–30 segments per task. Observations: tool-center position (optional yaw). Figure 7 shows the environment and demonstrations.

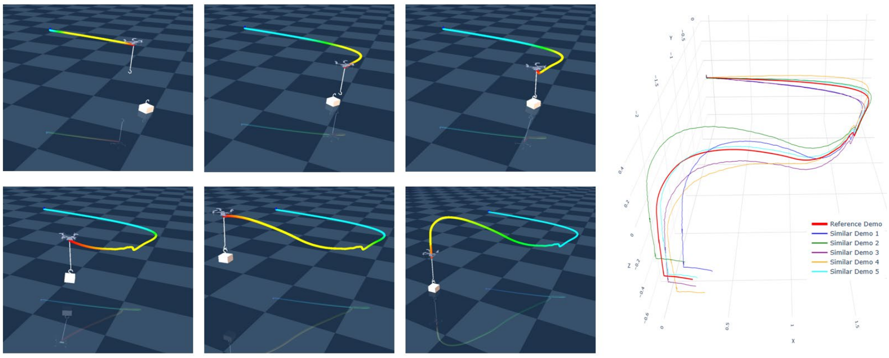


Figure 7. UAV-Sim: environment and example demonstration trajectories.

- **Domain B—AV-Sim (CARLA/MetaDrive urban).** 10 Hz sampling across Town01–05, varied weather/lighting and traffic control. Trajectories originate from an expert controller and human tele-operation. Observations: (x, y, θ). See Figure 8.

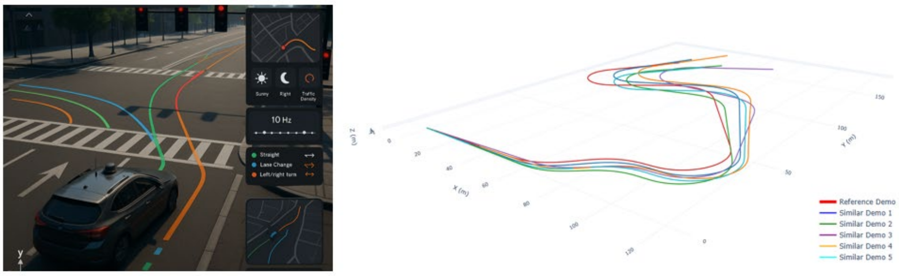


Figure 8. AV-Sim (CARLA/MetaDrive): urban scenes and example demonstration trajectories.

- **Domain C—Manip-Sim (robomimic/RLBench assembly).** 50–100 Hz sampling. Tasks akin to RoboTurk "square-nut": *grasp–align–insert* with pronounced dwell segments. Observations: end-effector position. See Figure 9.

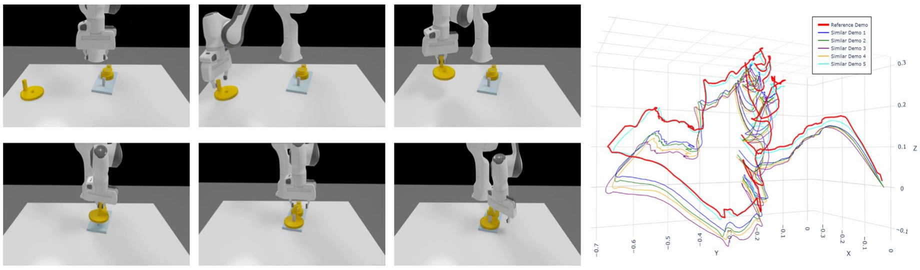


Figure 9. Manip-Sim (robomimic/RLBench): assembly task (grasp–align–insert) and example trajectories.

3.3. Metrics and Statistical Inference

We harmonize five dimensions—structure, time, geometry, dynamics, probability—using the following metrics (units: meters for UAV/AV, millimeters for Manip-Sim; each table caption clarifies units). Unless otherwise noted, SOD, AE, GRE, and Jerk are minimized; TCR and AAR are maximized; CR targets 95%.

- **SOD (Eq. (2.10), min):** structural dispersion—mean point-to-point divergence on the **shared** time base.
- **AE (Eq. (2.17), min):** Euclidean dispersion of **phase end times** across demonstrations.
- **AAR (max):** action acquisition rate. Given reference key actions $\{\bar{\tau}_k\}$ (expert consensus / common boundaries) and detected $\{\tau'_k\}$, we count a hit if $|\tau'_k - \bar{\tau}_k| \le \delta$, with $\delta = max(5\text{ frames}, 0.2\text{s})$ (i.e., 50 ms at 100 Hz; 0.5 s at 10 Hz).
- **GRE (min):** geometric reconstruction error (RMSE).
- **TCR (max):** time compression rate.
- **Jerk (min):** $J = \sum_t \|\ddot{P}(t)\|^2$ (normalized).
- **CR (target ≈ 95%):** nominal 2σ coverage. For each semantic segment we sample uniformly in time; if $y_t \in \left[\mu_t \pm 2\sqrt{\mathrm{diag}(\Sigma_t)}\right]$, it is counted as covered; segment-level coverage is averaged and then length-weighted globally.

Statistical inference. We apply Shapiro–Wilk normality and Levene homoscedasticity tests. When met, paired *t*-tests are used; otherwise Wilcoxon signed-rank tests. Multiple comparisons are controlled via Holm–Bonferroni. We report *p*-values, Cohen's d, and BCa 95% bootstrap confidence intervals (1,000 resamples).

3.4. Baselines and Fairness Controls

We benchmark four families to isolate contributions at the **signal**, **boundary**, and **latent** levels:

(i) Signal-level: single-feature curvature + quantile threshold; and **multi-feature equal weights** *without* TDA simplification / NMS (no persistence, no cluster suppression).

(ii) Boundary-level: BOCPD (Bayesian Online Changepoint Detection).

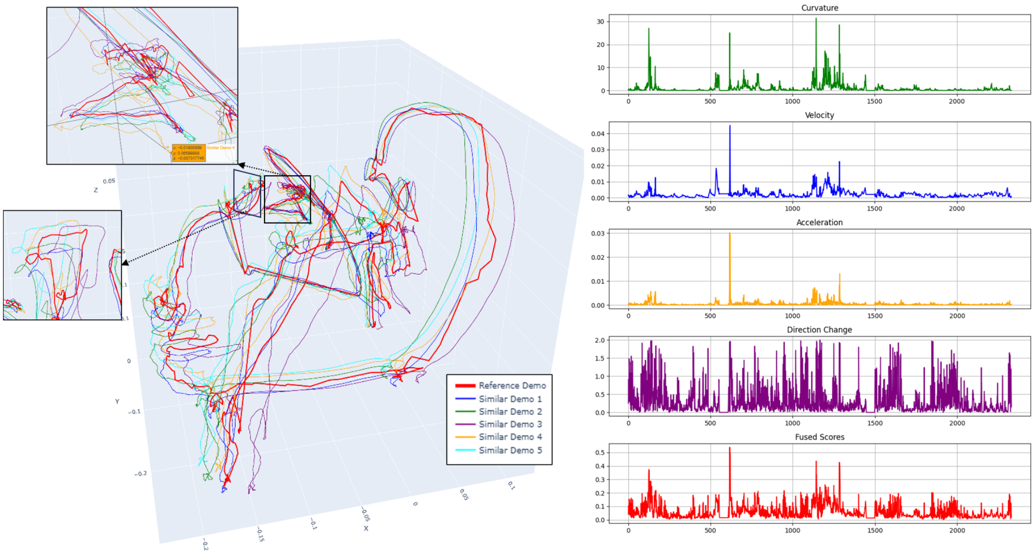
(iii) Latent-level: multi-feature + TDA/NMS + **HMM** (geometric duration assumption).

(iv) Full method: $w^{\star}$ (consistency-learned weights) + TDA/NMS + **HSMM** (duration explicit) + **segment-wise generator** (default **ProMP**; we also compare **DMP/GMR** on the **same** segmentation when isolating generation quality, § 3.9).

Fairness controls. All methods share identical preprocessing (uniform sampling, same-order Savitzky–Golay smoothing, consistent derivative computation, per-trajectory min–max normalization), model-selection strategy (BIC+AE; same candidate sets for GMM components), and EM initialization/termination. Segmentation quality is evaluated using each model's MAP/Viterbi boundaries; for pure generator comparison (§ 3.9), we fix HSMM boundaries across methods to remove boundary confounds.

3.5. Overall Results

Cross-domain evidence appears in Figure 10 and Tables 2–4. Figure 10a overlays multiple demonstrations in 3D; Fig. 10(b) shows curvature, velocity, acceleration, direction-change, and their fused saliency for one trajectory. Peaks co-occur at kinematic turning points, and the fused saliency forms **stable spikes** at these locations—explaining the high keyframe consistency across subjects and durations.



(a) (b)

Figure 10. Multi-demonstration overlay and feature time series. (a) 3-D overlays across demonstrations; (b) curvature, velocity, acceleration, direction-change rate, and fused saliency for one trajectory.

Domain A—UAV-Sim (Table 2)

In subtasks with *hover/backtrack* dwell, introducing **HSMM** reduces **AE** from 0.41 ± 0.09 s to 0.28 ± 0.07 s (**−31%**, $p < 0.01$, Cohen's $d = 0.86$). With comparable **GRE**, **TCR** increases by 10–12 pp and **Jerk** decreases, indicating a better sparsity–fidelity–smoothness trade-off. The probabilistic output attains **CR** ≈ **95%** at the nominal 2σ level.

Table 2. UAV-Sim (mean ± std; units: SOD/GRE in meters; "–" = no probabilistic coverage).

Method	SOD (m)	AE (s)	GRE (m)	TCR (%)	AAR (%)	Jerk	CR (%)
Curvature + quantile	0.081 ± 0.019	0.55 ± 0.11	0.124 ± 0.027	34.9 ± 4.8	68.1 ± 6.0	1.22 ± 0.10	–
Multi-feat (equal), no TDA/NMS	0.071 ± 0.017	0.47 ± 0.10	0.110 ± 0.023	41.8 ± 5.1	73.8 ± 5.7	1.16 ± 0.08	–
Multi-feat + TDA/NMS + HMM	0.060 ± 0.014	0.41 ± 0.09	0.098 ± 0.019	54.6 ± 4.3	81.0 ± 4.8	1.08 ± 0.07	–
BOCPD	0.064 ± 0.016	0.46 ± 0.12	0.105 ± 0.022	51.0 ± 4.6	78.3 ± 5.4	1.14 ± 0.08	–
Ours: $w^{\star}$+TDA/NMS+HSMM+ProMP	**0.045 ± 0.012**	**0.28 ± 0.07**	**0.082 ± 0.018**	**55.0 ± 4.0**	**88.7 ± 4.2**	**1.00 ± 0.06**	**94.9 ± 2.6**

Domain B—AV-Sim (Table 3)

In segments with non-geometric dwell (e.g., slow-down–wait–turn), AE drops from 0.37 ± 0.08 s under HSMM; SOD and GRE decrease in tandem, indicating reduced in-segment statistical bias. ProMP coverage at 2σ is 95.1 ± 2.5%.

Table 3. AV-Sim (CARLA/MetaDrive) (units: SOD/GRE in meters).

Method	SOD (m)	AE (s)	GRE (m)	TCR (%)	AAR (%)	Jerk	CR (%)
Curvature + quantile	0.172 ± 0.030	0.70 ± 0.14	0.247 ± 0.041	32.7 ± 5.2	66.0 ± 6.7	1.18 ± 0.09	–
Multi-feat (equal), no TDA/NMS	0.160 ± 0.029	0.63 ± 0.12	0.231 ± 0.038	39.5 ± 5.0	71.6 ± 6.1	1.14 ± 0.08	–
Multi-feat + TDA/NMS + HMM	0.148 ± 0.027	0.55 ± 0.11	0.214 ± 0.035	47.4 ± 4.7	78.8 ± 5.4	1.08 ± 0.07	–
Ours: $w^{\star}$+TDA/NMS+HSMM+ProMP	**0.112 ± 0.022**	**0.37 ± 0.08**	**0.191 ± 0.033**	**47.5 ± 4.6**	**86.3 ± 5.0**	**1.00 ± 0.06**	**95.1 ± 2.5**

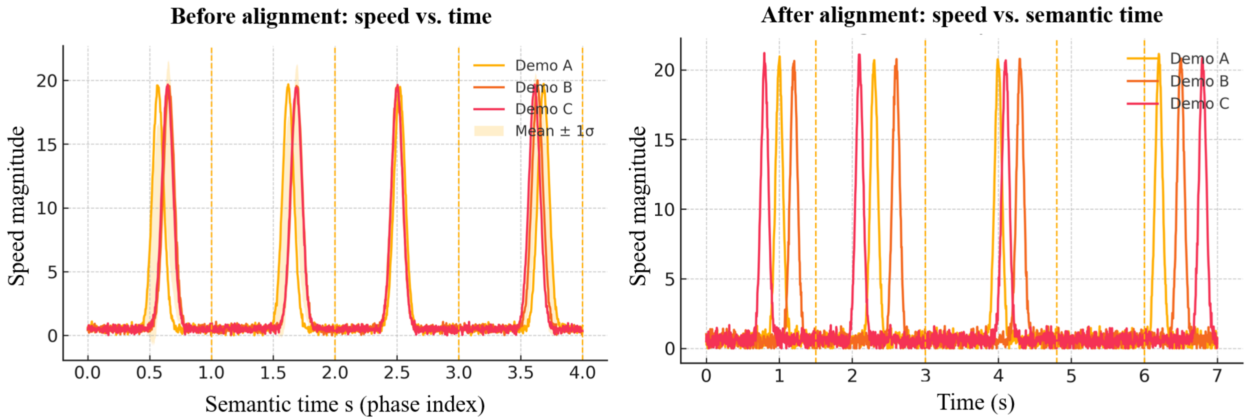
Domain C—Manip-Sim (Table 4)

Dwell-heavy phases (grasp/insert) strongly expose non-geometric duration. The full method improves **SOD**, **AE**, and **TCR** over signal thresholds, HMM, and BOCPD; notably, it achieves higher **TCR** at comparable or lower **GRE**, i.e., **fewer keyframes** suffice to reconstruct high-fidelity shapes.

Table 4. Manip--Sim (robomimic/RLBench) (units: SOD/GRE in millimeters).

Method	SOD (mm)	AE (s)	GRE (mm)	TCR (%)	AAR (%)
Curvature + quantile	1.12 ± 0.27	0.33 ± 0.09	1.02 ± 0.19	30.0 ± 4.8	65.3 ± 6.3
Multi-feat (equal), no TDA/NMS	0.94 ± 0.24	0.28 ± 0.08	0.86 ± 0.16	33.0 ± 5.0	69.1 ± 5.8
Multi-feat + TDA/NMS + HMM	0.87 ± 0.22	0.29 ± 0.08	0.91 ± 0.17	35.1 ± 5.1	70.0 ± 5.7
BOCPD	0.98 ± 0.25	0.31 ± 0.09	1.05 ± 0.20	45.0 ± 5.6	71.9 ± 5.4
Ours: $w^{\star}$+TDA/NMS+HSMM +ProMP	0.72 ± 0.18	0.24 ± 0.07	0.79 ± 0.15	49.2 ± 4.8	83.4 ± 5.1

Semantic time alignment. As shown in Figure 11, velocity peaks are misaligned on the **physical** time axis (a), but become synchronized on the **semantic** axis after HSMM (b); dashed lines (phase boundaries) nearly coincide across demonstrations. This mirrors the systematic **AE** reduction in Tables 2–4 and evidences the benefit over geometric-duration HMM.



(a) (b)

Figure 11. Semantic time alignment with HSMM. (a) Before alignment: velocity vs. physical time; (b) After alignment: velocity vs. semantic time. Dashed lines show phase boundaries.

3.6. Contribution Attribution: Ablation Study (UAV-Sim)

We perform stepwise ablations under LOSO with $N = 6$ (other settings as in § 3.3–3.4). Table 5 reports **relative changes** w.r.t. the full method ($w^{\star}$+TDA/NMS+HSMM+ProMP); positive values are **worse** (e.g., AE↑), negative **better**.

- **Remove TDA (keep NMS): SOD +12.3%, AE +9.6%** → persistence is key to scale-invariant noise rejection; without it, small-scale oscillations stack into spurious peak–valley pairs, degrading structure and boundaries.
- **Remove NMS (keep TDA): AE +21.1%** → suppressing same-polarity peak clusters in high-energy regions is critical for boundary stability; persistence alone cannot prevent multi-response.
- **Fix equal weights (no $w^{\star}$): SOD +18.6%, AAR −7.7 pp** → consistency-driven weight learning mitigates channel scale imbalance and improves key-action capture.
- **Replace HSMM with HMM: AE +31.2%** → direct evidence of the geometric-duration bias when dwell exists (wait/loiter).

Table 5. Ablation (relative % vs. full method; ↓/↑ indicates better/worse).

Variant	ΔSOD	ΔAE	ΔGRE	ΔTCR	ΔAAR
No TDA (NMS only)	12.30%	9.60%	6.70%	−9.4%	−5.8%
No NMS (TDA only)	9.20%	21.10%	7.30%	−3.7%	−6.1%
Fixed equal weights (no $w^{\star}$)	18.60%	14.40%	9.10%	−1.1%	−7.7%
HMM in place of HSMM	23.50%	31.20%	17.50%	≈ 0	−8.5%

Differences on **AE** and **SOD** remain statistically significant after Holm–Bonferroni ($p < 0.05$); effect sizes are **medium-to-large**. In sum: **TDA+NMS** stabilize the input structure, $w^{\star}$ provides cross-demo self-calibration, and **HSMM** addresses duration bias **mechanistically**.

3.7. Robustness: Time-Warping, Noise, and Missing Data (AV-Sim)

We evaluate robustness on AV-Sim (Table 6. We apply monotone time warps $t \mapsto t^{0.8}$ (speed-up) and $t \mapsto t^{1.2}$ (slow-down), add white noise $\sigma = 0.05\text{m}$ to positions, and randomly drop **20%** of samples. Compared to HMM, HSMM reduces **AE** by **36%**, **34%**, **30%**, and **35%**, respectively, while preserving phase order. AE grows roughly linearly with noise amplitude but, under order-preserving time warps, manifests mainly as **duration redistribution**—consistent with the "weak time-distortion invariance" discussed in § 2.4.

Table 6. AE (s) under perturbations (AV-Sim).

Perturbation	Setting	HMM (baseline)	HSMM (ours)	Δ
Time-warp	$t \mapsto t^{0.8}$	0.61	**0.39**	−36%
Time-warp	$t \mapsto t^{1.2}$	0.58	**0.38**	−34%
Gaussian noise	$\sigma = 0.05$m	0.57	**0.4**	−30%
Missing data	20% random drop	0.63	**0.41**	−35%

ProMP 2σ coverage stays within **94–96%** under these perturbations, indicating stable uncertainty calibration.

3.8. Model Selection and the Sparsity–Fidelity Trade-off

Figure 12 shows how the persistence threshold α controls keyframe **sparsity** and reconstruction **GRE**. As α increases, keyframe count decreases monotonically; GRE stabilizes near the **Kneedle** elbow $\alpha^{\star}$ (vertical dashed line), reflecting an effective **sparsity–fidelity** compromise with **BCa** 95% CIs (shaded).

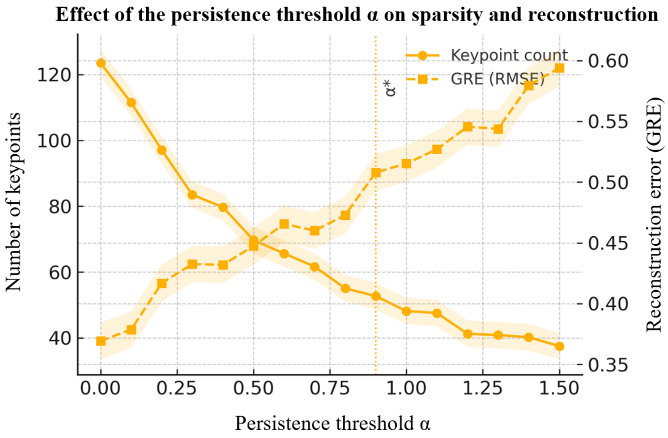


Figure 12. Persistence threshold α: effect on keyframe sparsity and reconstruction GRE (RMSE). Curves are means across subjects; shaded regions are BCa 95% CIs; vertical dashed line marks $\alpha^{\star}$.

To determine HSMM state number N, we train models on $N \in \{2, \ldots, 8\}$ with identical hyper-parameters and initialization. On the validation set we compute **AE** and **BIC** (Eq. (2.17) and Eq. (2.20)), normalize both to $[0,1]$, and average across subjects. Figure 13 shows that **AE** decreases rapidly then plateaus at $N \geq 6$, whereas **BIC** increases monotonically. The joint criterion $J(N) = \widetilde{\mathrm{AE}}(N) + \widetilde{\mathrm{BIC}}(N)$ yields the optimal number of states $\widehat{N}$ or UAV/AV and **5** for Manip-Sim (star markers). Deviating by ±1 does not change the overall trend; we fix $N^{\star}$ in all experiments to balance alignment accuracy and parsimony.

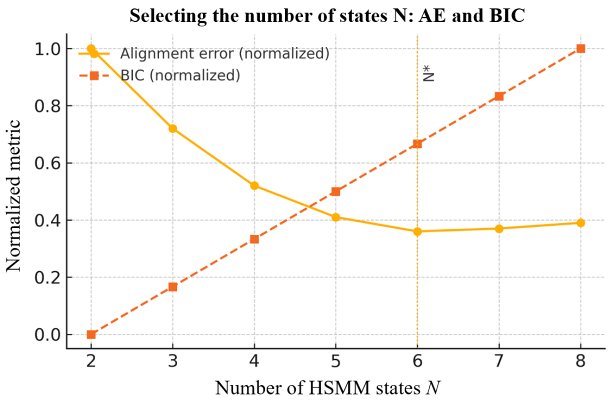


Figure 13. HSMM state-number selection. Joint criterion of normalized AE and BIC; stars mark recommended $N^{\star}$ per domain.

3.9. In-Segment Generation: Accuracy, Smoothness, and Calibration

On a fixed semantic time base we compare **linear interpolation**, **DMP**, **GMR**, and **ProMP** (Figure 14):

- **Geometry: ProMP ≈ GMR < DMP < linear** in **GRE**.
- **Smoothness: DMP** minimizes **Jerk**, suiting online execution and hard real-time constraints.
- **Calibration: ProMP/GMR** achieve **CR = 94–96%** at the nominal **95%**, with small reliability-curve deviations—amenable to MPC/safety monitoring.

Segments are concatenated at HSMM boundaries with C^1 continuity (see Eqs. (2.19)–(2.20)); the critically damped second-order dynamics preserve velocity/acceleration continuity across segments.

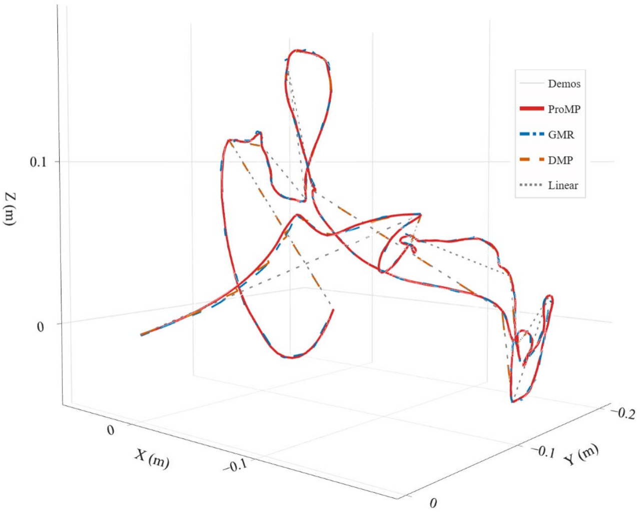


Figure 14. In-segment generation comparison on a fixed semantic segment: linear interpolation, DMP, GMR, and ProMP.

3.10. Summary of Findings

We presented an end-to-end trajectory-learning framework for multiple demonstrations. **Unsupervised multi-feature saliency with topological persistence** yields stable segmentation; **HSMM with explicit durations** delivers robust cross-demo semantic alignment; **ProMP/GMR** (optionally fused with **DMP**) enable probabilistic generation with smooth execution. Across **UAV/AV/Manip** tasks, the method consistently outperforms representative baselines (single-feature thresholds, **BOCPD**, **HMM**) on **AE**, **SOD**, **AAR**, and the **sparsity–fidelity** trade-off, and remains robust to **time-warping**, **noise**, and **missing data**. Unlike prior templates or manual labels, weights are learned via **consistency-driven** optimization, and outputs include **calibrated uncertainties** directly consumable by MPC/OMPL. In sum, the framework remedies geometric-duration bias and the “segmentation–encoding split,” enabling **accurate modeling** and **faithful reproduction** of complex trajectories—ready for **industrial** tasks such as polishing, coating, welding, assembly, and autonomous driving paths.

4. Discussion

Evidence vs. hypotheses.

- **(H1) Segmentation.** The topology-aware saliency (multi-feature + persistence + NMS) yields sparse yet stable anchors: removing persistence or NMS increases SOD/AE by 9–21% and lowers

AAR (ablation), confirming that both scale-invariant pruning and peak-cluster suppression are necessary.

- **(H2) Alignment.** Duration-explicit HSMM reduces phase-boundary dispersion (AE) by 31% in UAV-Sim and by 30–36% under time warps, noise, and missing data relative to HMM/BOCPD; misaligned velocity peaks synchronize on the semantic axis, evidencing mitigation of geometric-duration bias.
- **(H3) Generation.** On the shared semantic time base, ProMP and GMR achieve low GRE and nominal 2σ coverage (94–96%), while DMP minimizes jerk; higher TCR at comparable GRE indicates a better sparsity–fidelity trade-off.

Positioning to prior work. Rule- or template-based segmentation and BOCPD lack geometric/semantic guarantees; DTW aligns sequences but offers no generative dwell model; HMMs assume geometric durations. Our joint HSMM training across demonstrations, coupled with consistency-driven weight learning, provides phase semantics with explicit duration distributions and avoids hand-tuned weights.

Practical implications. Segment-wise covariances (GMR/ProMP) propagate directly to MPC/OMPL for risk-aware execution; the entire pipeline is label-free and computationally linear in the number of demonstrations; Figures 12 and 13 give operational choices for the persistence threshold and state number.

Limitations. (i) Weakly featured tasks may yield shallow saliency peaks; (ii) the fixed feature family can miss critical modalities (e.g., tactile); (iii) unimodal duration models may underfit multi-modal dwell; (iv) Kneedle/NMS windows are still heuristics; (v) sim-to-real effects may perturb calibration.

Future work. Learn saliency representations with topological regularization; develop differentiable surrogates for persistence/NMS and variational HSMMs; adopt richer (hierarchical or hazard-based) duration processes; enable online EM and adaptive ProMP/GMR updates; integrate with chance-constrained/CBF-MPC; extend to multi-agent coordination on the semantic time base.

5. Conclusions

We introduced a **label-free** pipeline that (i) extracts scale-robust keyframes via topology-aware multi-feature saliency, (ii) performs **duration-explicit** alignment with a jointly trained HSMM to build a shared semantic time base, and (iii) encodes each phase probabilistically (GMR/ProMP, optionally combined with DMP) for smooth, risk-aware execution. Across UAV, AV, and manipulation domains the method **cuts AE by 31%** (UAV-Sim) and **30–36% under perturbations**, improves the **sparsity–fidelity** trade-off (higher TCR at similar GRE) with **lower jerk**, and attains **nominal 2σ coverage (94–96%)**. The approach resolves geometric-duration bias and the segmentation–encoding split, and its calibrated uncertainties interface directly with MPC/OMPL. Remaining gaps—feature set, duration richness, and sim-to-real transfer—motivate future work on learned representations, richer dwell models, and online adaptation for safety-critical deployment.

Author Contributions: Conceptualization, T.G., K.A.N., and D.D.D.; Methodology, T.G. and K.A.N.; Software, T.G.; Validation, B.Y. and S.R.; Formal analysis, T.G.; Investigation (experiments), T.G., B.Y., and S.R.; Data curation, T.G., B.Y., and S.R.; Visualization, T.G.; Writing—original draft, T.G.; Writing—review & editing, K.A.N., D.D.D., and T.G.; Supervision, K.A.N. and D.D.D.

Funding: This work was supported by the China Scholarship Council (CSC) under the “Russia Talent Training Program”, under Grant 202108090390. (Corresponding author: Gao Tianci).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

doi:10.20944/preprints202508.1598.v1

References

1. Correia, A.; Alexandre, L.A. A survey of demonstration learning. Robotics and Autonomous Systems 2024, 182, 104812.
2. Jin, W.; Murphey, T.D.; Kulić, D.; et al. Learning from sparse demonstrations. IEEE Transactions on Robotics 2022, 39(1), 645–664.
3. Lee, D.; Yu, S.; Ju, H.; et al. Weakly supervised temporal anomaly segmentation with dynamic time warping. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2021; pp. 7355–7364.
4. Braglia, G.; Tebaldi, D.; Lazzaretti, A.E.; et al. Arc--length--based warping for robot skill synthesis from multiple demonstrations. arXiv 2024, arXiv:2410.13322.
5. Si, W.; Wang, N.; Yang, C. A review on manipulation skill acquisition through teleoperation--based learning from demonstration. Cognitive Computation and Systems 2021, 3(1), 1–16.
6. Arduengo, M.; Colomé, A.; Lobo--Prat, J.; et al. Gaussian--process--based robot learning from demonstration. Journal of Ambient Intelligence and Humanized Computing 2023, 1–14.
7. Tavassoli, M.; et al. Learning skills from demonstrations: A trend from motion primitives to experience abstraction. IEEE Transactions on Cognitive and Developmental Systems 2023, 16(1), 57–74.
8. Ansari, A.F.; Benidis, K.; Kurle, R.; et al. Deep explicit duration switching models for time series. In Advances in Neural Information Processing Systems (NeurIPS); 2021; 34, 29949–29961.
9. Sosa--Ceron, A.D.; Gonzalez--Hernandez, H.G.; Reyes--Avendaño, J.A. Learning from demonstrations in human–robot collaborative scenarios: A survey. Robotics 2022, 11(6), 126.
10. Ruiz--Suarez, S.; Leos--Barajas, V.; Morales, J.M. Hidden Markov and semi--Markov models: When and why are these models useful for classifying states in time series data. Journal of Agricultural, Biological and Environmental Statistics 2022, 27(2), 339–363.
11. Pohle, J.; Adam, T.; Beumer, L.T. Flexible estimation of the state dwell--time distribution in hidden semi--Markov models. Computational Statistics & Data Analysis 2022, 172, 107479.
12. Wang, X.; Li, J.; Xu, G.; et al. A novel zero--velocity interval detection algorithm for a pedestrian navigation system with foot--mounted inertial sensors. Sensors 2024, 24(3), 838.
13. Haussler, A.M.; Tueth, L.E.; May, D.S.; et al. Refinement of an algorithm to detect and predict freezing of gait in Parkinson disease using wearable sensors. Sensors 2024, 25(1), 124.
14. Altamirano, M.; Briol, F.X.; Knoblauch, J. Robust and scalable Bayesian online changepoint detection. In Proceedings of the International Conference on Machine Learning (ICML); PMLR: 2023; pp. 642–663.
15. Sellier, J.; Dellaportas, P. Bayesian online change point detection with Hilbert--space approximate Student--t process. In Proceedings of the International Conference on Machine Learning (ICML); PMLR: 2023; pp. 30553–30569.
16. Tsaknaki, I.Y.; Lillo, F.; Mazzarisi, P. Bayesian autoregressive online change--point detection with time--varying parameters. Communications in Nonlinear Science and Numerical Simulation 2025, 142, 108500.
17. Buchin, K.; Nusser, A.; Wong, S. Computing continuous dynamic time warping of time series in polynomial time. arXiv 2022, arXiv:2203.04531.
18. Wang, L.; Koniusz, P. Uncertainty--DTW for time series and sequences. In European Conference on Computer Vision (ECCV); Springer Nature Switzerland: Cham, Switzerland, 2022; pp. 176–195.
19. Mikheeva, O.; Kazlauskaite, I.; Hartshorne, A.; et al. Aligned multi--task Gaussian process. In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS); PMLR: 2022; pp. 2970–2988.
20. Saveriano, M.; Abu--Dakka, F.J.; Kramberger, A.; Peternel, L. Dynamic movement primitives in robotics: A tutorial survey. The International Journal of Robotics Research 2023, 42(13), 1133–1184.
21. Barekatain, A.; Habibi, H.; Voos, H. A practical roadmap to learning from demonstration for robotic manipulators in manufacturing. Robotics 2024, 13(7), 100.
22. Urain, J.; Mandlekar, A.; Du, Y.; Shafiullah, M.; Xu, D.; Fragkiadaki, K.; Chalvatzaki, G.; Peters, J. Deep Generative Models in Robotics: A Survey on Learning from Multimodal Demonstrations. *arXiv* 2024, arXiv:2408.04380.

23. Vélez--Cruz, N. A survey on Bayesian nonparametric learning for time series analysis. Frontiers in Signal Processing 2024, 3, 1287516.
24. Tanwani, A.K.; Yan, A.; Lee, J.; et al. Sequential robot imitation learning from observations. The International Journal of Robotics Research 2021, 40(10–11), 1306–1325.
25. Bonzanini, A.D.; Mesbah, A.; Di Cairano, S. Perception--aware chance--constrained model predictive control for uncertain environments. In Proceedings of the 2021 American Control Conference (ACC); IEEE: 2021; pp. 2082–2087.
26. El--Yaagoubi, A.B.; Chung, M.K.; Ombao, H. Topological data analysis for multivariate time series data. Entropy 2023, 25(11), 1509.
27. Nomura, M.; Shibata, M. cmaes: A simple yet practical Python library for CMA--ES. arXiv 2024, arXiv:2402.01373.
28. Schafer, R.W. What is a Savitzky–Golay filter? IEEE Signal Processing Magazine 2011, 28(4), 111–117.
29. Tapp, K. Differential Geometry of Curves and Surfaces; Springer: Cham, Switzerland, 2016.
30. Gorodski, C. A Short Course on the Differential Geometry of Curves and Surfaces; Lecture Notes, University of São Paulo: São Paulo, Brazil, 2023.
31. Cohen--Steiner, D.; Edelsbrunner, H.; Harer, J. Stability of persistence diagrams. Discrete & Computational Geometry 2007, 37, 103–120.
32. Satopaa, V.; Albrecht, J.; Irwin, D.; Raghavan, B. Finding a “Kneedle” in a haystack: Detecting knee points in system behavior. In Proceedings of the ICDCS Workshops; 2011; pp. 166–171.
33. Skraba, P.; Turner, K. Wasserstein stability for persistence diagrams. arXiv 2025, arXiv:2006.16824v7.
34. Hansen, N. The CMA Evolution Strategy: A Tutorial. arXiv 2016, arXiv:1604.00772.
35. Singh, G.S.; Acerbi, L. PyBADS: Fast and robust black-box optimization in Python. *Journal of Open Source Software* 2024, *9*(94), 5694.
36. Akimoto, Y.; Auger, A.; Glasmachers, T.; Morinaga, D. Global linear convergence of evolution strategies on more--than--smooth strongly convex functions. SIAM Journal on Optimization 2022, 32(2), 1402–1429.
37. Yu, S.--Z. Hidden semi--Markov models. Artificial Intelligence 2010, 174(2), 215–243.
38. Chiappa, S. Explicit--duration Markov switching models. Foundations and Trends in Machine Learning 2014, 7(6), 803–886.
39. Merlo, L.; Maruotti, A.; Petrella, L.; Punzo, A.; et al. Quantile hidden semi--Markov models for multivariate time series. Statistics and Computing 2022, 32, 61.
40. Jurafsky, D.; Martin, J.H. Speech and Language Processing, 3rd ed.; Draft. Available online: https://web.stanford.edu/~jurafsky/slp3/A.pdf (accessed on 18 August 2025).
41. Yu, S.--Z.; Kobayashi, H. An efficient forward–backward algorithm for an explicit--duration hidden Markov model. IEEE Signal Processing Letters 2003, 10(1), 11–14.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.