PIPF: Physics-Informed Path Integral Filter-Based Video Text Tracking

[1]Shubham Das, [1]Tamanas Sahana, [2]Shivakumara Palaiahnakote, [1]Umapada Pal, [2]Apostolos Antonacopoulos
[1]Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, India.
mdas14800@gmail.com, tamanassahana@gmail.com, umapada@isical.ac.in
[2]School of Science, Engineering and Environment, University of Salford, United Kingdom
S.Palaiahnakote@salford.ac.uk, A.Antonacopoulos@primaresearch.org

Abstract. Tracking multiple non-linear text instances in videos is a complex challenge, and most existing methods rely on supervised learning for text tracking. Towards multiple non-linear video text tracking, in this paper, we have proposed a novel framework using Physics-Informed Path Integral Filter (PIPF), a training-free framework that enables zero-shot video text tracking by grounding motion in classical mechanics. PIPF formulates motion forecasting as a stochastic optimal control problem, solving the Feynman-Kac path integral via Monte Carlo sampling to minimize a thermodynamic action functional derived from the Hamilton-Jacobi-Bellman equation. This explicitly encodes Lagrangian dynamics, bounded turning, and geometric constraints, isolating motion priors from visual representations for robust, interpretable inference without any video-specific labels or learning, which is suitable for multiple non-linear video text tracking. A key novelty is PIPF's plug-and-play architecture, which seamlessly converts any frozen static text detector (e.g., CRAFT or DBNet++) and foundation model embedder (e.g., CLIP or DINOv2) into a competitive video tracker, enhancing performance in existing architectures. By combining these with PIPF, the framework handles non-linear motions, occlusions, and camera shakes effectively. Evaluated on ICDAR 2015 Video and ICDAR 2023 DSText benchmarks, despite of having zero video training, the proposed framework achieves comparable results with many supervised, outperforming a few older supervised methods as well. The main strength lies in PIPF's role as a scalable Data Engine, generating silver-standard trajectory annotations from raw videos at zero marginal cost. This offers a rigorous, instantly deployable alternative to annotation-intensive end-to-end supervision, with performance bounded only by the underlying detector in extreme occlusions. The source code will be made available at https://github.com/mdas14800-Shubham/PIPF-Paper-Repository.git.

Keywords: Video Text Tracking, Zero-Shot technique, Physics-Informed Filtering.

1 Introduction

Video text tracking involves localizing textual regions in video streams while preserving instance identity across video frames, which is a core capability for autonomous

driving, augmented reality, video retrieval, document understanding, and real-time content analysis. Compared to general multi-object tracking, text tracking instances pose extreme challenges like arbitrarily high aspect ratios, dense overlapping layouts, severe motion blur, unconstrained orientations, and rapid appearance changes driven by camera ego-motion rather than object articulation, etc. Thus, video text tracking is a challenging task.

Recent advances in video text tracking have predominantly relied on large-scale end-to-end supervised learning [1-6]. Modern frameworks jointly train text-specific detectors, re-identification networks, and spatio-temporal motion models on tens of thousands of meticulously annotated video frames. While undeniably effective when abundant labelled video data is available, this paradigm suffers from three critical limitations: (i) prohibitive annotation cost (precise polygons on every frame), (ii) poor generalization to unseen motion patterns, blur distributions, or camera types, and (iii) architectural complexity that hinders deployment on edge devices.

In parallel, the vision-language community has shown that powerful visual and semantic understanding can be obtained without task-specific labels by leveraging large-scale image-text pretraining (CLIP [7], Florence [8]) and static-image specialist models (CRAFT [9], DBNet++ [10]). A natural question, therefore, arises: Can robust video text tracking be achieved without any video-specific training whatsoever?

This work answers affirmatively. As shown in Fig. 1(a) and (b), the existing models follow the conventional approach that uses a large number of training samples. The proposed model does not require samples and labels for learning to determine trajectories. In addition, Fig. 2 shows that standard tracking models like Kalman motion and a frozen CLIP-Hungarian model [7] miss the target text ID, especially for 223 and 235 frames, due to non-linear, rapid text movements. However, the proposed model detects the target text ID accurately, irrespective of text motions in the video. This is because, in contrast to the existing models that learn motion from linear motion models, the proposed model directly imposes classical mechanics (inertia, bounded turning, image boundary avoidance) via a stochastic optimal control formulation [11, 12] instead of training re-identification networks. In summary, the proposed model relies on the semantic invariance already captured by frozen CLIP [7] or DINOv2 [13] embeddings. Therefore, our contributions are as follows.

(i) A novel Physics-Informed Path Integral Filter (PIPF) that formulates motion forecasting as a Feynman–Kac path integral over physically constrained trajectories is proposed for video text tracking. (ii) A plug-and-play proposed framework that instantly converts any frozen static text detector (CRAFT [9], DBNet++ [10]) and any foundation-model embedder (CLIP [7], DINOv2 [13]) into a competitive video tracker without architectural changes (iii) Unlike other methods, the proposed method can handle multiple non-linear trajectory text instances in video (iv) To the best of our knowledge, this is the first work where PIPF concept is used for video text tracking purpose.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 details the proposed methodology. Section 4 presents extensive experiments and ablation studies. We discuss limitations and conclusions in Sections 5 and 6, respectively.

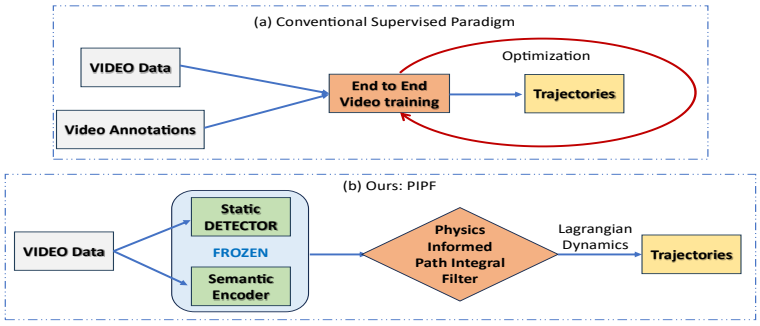

Fig. 1. A visual comparison of our pipeline with that of conventional methods.

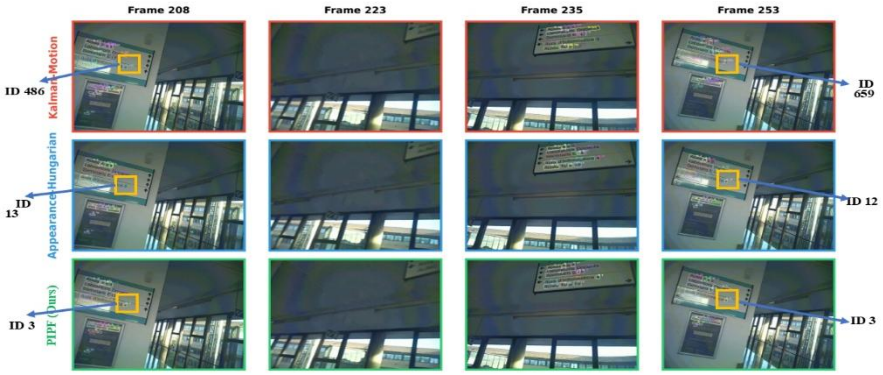

Fig. 2. Qualitative comparison of our method with baselines. Frames 223 and 235 are going through rapid motion and occlusion for target ID. Our method PIPF (3rd row) successfully maintains the same id while other two (1st and 2nd rows) is unable to do it. This is further discussed in the Ablation study, Section 4.4. More video results can be found at link.

2 Literature Review

To gain an understanding of the current state of the approaches, we review here the methods related to text tracking, including both conventional and modern techniques.

2.1 Frozen Scene Text Detection

Robust single-frame detection is the fundamental prerequisite for tracking. Early regression-based methods like TextBoxes++ [14] and EAST [15] optimized anchor scales and pixel-level regression for speed but often struggled with curve distortions. Segmentation-based approaches such as PSENet [16], Differentiable Binarization (DB) [17], CRAFT [9], and DBNet++ [10] currently dominate thanks to their ability to handle curved and densely packed text. We treat these models as frozen priors without any video-specific fine-tuning. In summary, there are several text detection methods for addressing challenges of detection that are effective for single-frame or natural scene

images, but not for video. Therefore, the scope of the models is limited to a single frame.

2.2 Video Text Tracking: The Supervised Paradigm

Video Text Tracking has evolved from heuristic tracking-by-detection (SORT [18], DeepSORT [19]) to fully supervised end-to-end learning. Recent transformer-based methods [1, 2, 3] and parameter-efficient approaches [4] dominate ICDAR 2015 Video and ICDAR 2023 DSText leaderboards by jointly learning detection, recognition, and motion on large, annotated video collections [5, 6]. However, they remain data-hungry and generalize poorly across domains. There are models developed in the past for tracking text in video that use detection and tracking of text. It is noted from the past methods that none of the models are free-training and require labels for achieving the results. In addition, the scope of the methods is limited to tracking a single instance with linear motion. Therefore, the methods are not robust enough to handle real-time situations.

2.3 Zero-Shot and Foundation Models

Foundation-scale models enable perception without reliance on task-specific annotations. CLIP [7] demonstrated that learning from 400 million image-text pairs yields robust zero-shot semantics, a paradigm extended to spatio-temporal tasks by Florence [8]. In the self-supervised vision domain, DINOv2 [13] has set a new standard for producing discriminative visual features without text supervision. Our framework combines these frozen semantic (CLIP) and geometric (DINOv2) features with physical priors to bypass the need for task-specific training. The models used zero-shot learning to overcome the challenge of training on a large number of samples to make the method efficient and robust. However, these models are not tested on multiple text instances with non-linear motions in video.

2.4 Physics-Informed Tracking and Control

To handle non-linear motion without learning, we draw on stochastic optimal control. Kappen [11] and Theodorou et al. [12] established the Path Integral (PI) control framework, which solves non-linear control problems by weighting trajectories via a cost functional, a method popularized in robotics by CHOMP [20]. While physics-informed tracking has appeared in vision works like PhyOT [21], these often treat physics as a regularizer for neural networks. In contrast, PIPF employs the Path Integral as the core inference engine to forecast motion in a fully zero-shot manner. To model continuous state propagation accurately, underlying state dynamics are typically anchored to foundational stochastic parameters defined over structured motion horizons [26].

Overall, it is observed from the review on different approaches that none of the methods used the concept of PIPF for video text tracking to address the challenge of multiple instances with non-linear motion.

3 Proposed Methodology

As noted in the previous section, tracking multiple text instances with non-linear motions is an open challenge. To address such a challenge, we introduce a novel concept, Physics-Informed-Path-Integral-Filter (PIPF), with a fully zero-shot video text tracking framework that combines frozen detector-based polygon detection, semantic appearance modelling using a frozen encoder model. The PIPF is proposed for motion prediction derived from stochastic optimal control theory. We sequentially integrate robust spatial detection, semantic re-identification, and physics-based filtering to estimate trajectories $\boldsymbol{\mathcal{T}} = \{\boldsymbol{\tau_k}\}_{\boldsymbol{k=1}}^{\boldsymbol{N}}$ for a video sequence $\boldsymbol{\mathcal{V}} = \{\boldsymbol{I_1}, \boldsymbol{I_2}, \dots, \boldsymbol{I_T}\}$, where each trajectory $\boldsymbol{\tau_k} = \{(\boldsymbol{b_t}, \boldsymbol{e_t})\}_{\boldsymbol{t=t}_{\text{start}}}^{\boldsymbol{t}_{\text{end}}}$ encodes the spatial bounding polygon $\boldsymbol{b_t}$ and semantic embedding $\boldsymbol{e_t}$. The block diagram is presented in Fig. 3. The framework processes the Input Frame t by first employing a frozen Detector and Semantic Encoder to extract spatial polygons and normalized visual embedding. Simultaneously, the verified state from Output Frame $t-1$ (position, velocity, and reference embedding) is used to initialize Monte-Carlo Sampling, where N particle trajectories are generated via a second-order Stochastic Differential Equation (SDE). These trajectories undergo an Action Calculation that weights paths based on physical priors like kinetic smoothness and turn penalties to generate a weighted Prediction $(\hat{x}_{t|t-1})$. Finally, a Cost Matrix integrates these physical predictions with the current frame's semantic detections, utilizing the Hungarian Assignment for global association to produce the Output Frame t with maintained text identities.

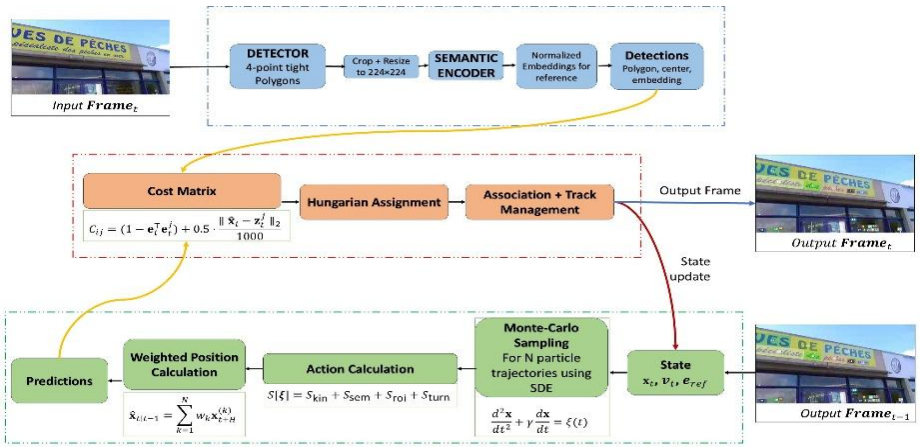

Fig. 3. Block diagram of the proposed model.

3.1 Scene Text Detection

Accurate spatial localization is critical for reliable embedding extraction and downstream tracking**.** We use CRAFT [9] pretrained on [22, 23] to predict pixel-wise

character region score $S_r$and affinity score S_a. The binary text instance map at pixel p is obtained via thresholding as defined in Equation (1).

$$M(p) = \mathbb{I}(S_r(p) > \tau_r \vee S_a(p) > \tau_a) \tag{1}$$

where τ_r(region threshold), τ_a(affinity threshold) for the CRAFT model, $S_r(p)$represents the probability that pixela p is a character center, $S_a(p)$indicates linkage between adjacent characters, and $\mathbb{I}[\cdot]$is the indicator function. $M(p) \in \{0,1\}$forms a binary mask for connected component analysis.

The final bounding polygons $b_t^i$are generated by computing the minimal enclosing hull for each connected component in $M(p)$, then fitted with a minimum-area rotated rectangle using OpenCV's minAreaRect to boxPoints, yielding exactly 4-point polygons $\{b_t^i \in \mathbb{R}^{4\times2}\}_{i=1}^{N_t}$as required by ICDAR evaluation protocol. Centroids $c_t^i = \frac{1}{4}\sum_{k=1}^{4} b_t^i(k)$ serve as positional states for tracking.

3.2 Zero-Shot semantic Embedding

To maintain identity coherence across appearance variations (e.g., blur, scale), we leverage the semantic robustness of frozen CLIP and DINOv2. For each detected 4-point polygon b_t^i, we extract its axis-aligned crop c_t^i, resize to 224×224, and compute a normalized visual embedding as per Eq. 2.

$$\mathbf{e}_t^i = \frac{\Phi_{\text{vis}}(c_t^i)}{\| \Phi_{\text{vis}}(c_t^i) \|_2 + \epsilon} \in \mathbb{S}^{512} \tag{2}$$

where Φ_{vis} is the frozen CLIP ViT-B/32 image encoder, $\mathbb{S}^{512}$denotes the 512-dimensional unit hypersphere, and $\epsilon = 10^{-12}$prevents division by zero. Active tracks maintain an exponentially smoothed reference embedding with smoothing constant α_e as per Eq. 3.

$$\mathbf{e}_{\text{ref}} \leftarrow (1-\alpha_e)\mathbf{e}_{\text{ref}} + \alpha_e \mathbf{e}_t, where\ \alpha_e = 0.2 \tag{3}$$

$$\text{Sim}(\mathbf{e}_i, \mathbf{e}_j) = \mathbf{e}_i^\top \mathbf{e}_j \in [-1,1] \tag{4}$$

Cosine similarity is used throughout for association, shown in Eq. 4. Figure 4 shows the strength of CLIP to form tight similar embeddings for Text Instance patches.

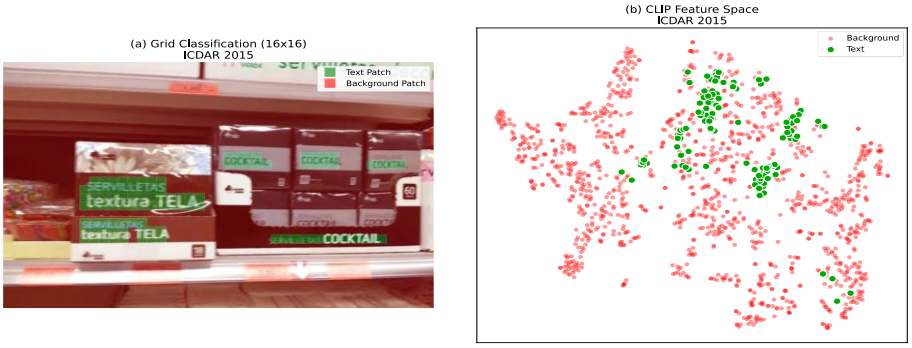

Fig. 4. Shows t-SNE visualization of CLIP embeddings for a single frame. Identical text content forms tight clusters despite viewpoint and scale changes.

3.3 Physics-Informed Path Integral Filter (PIPF)

Standard motion filters like Kalman assume linear motion with Gaussian noise, which fails under videos having non-linearities like camera shake. We instead model each track's centroid $\mathbf{x}_t \in \mathbb{R}^2$as a controlled stochastic process, solving the forward prediction via path integral control, a principled framework from stochastic optimal control that marginalizes over physically plausible futures by using the Law of Least Action. We use a fixed prediction horizon of $T = 5$ frames (approximately 0.2 seconds at 25 FPS), which we found sufficient to capture sharp turns and camera motion while keeping computation lightweight.

State Dynamics. The state evolves under second-order double-integrator with mean reverting accelerating noise, discretized from the continuous time Stochastic Differential Equation (SDE) as

$$d\mathbf{v} = \mathbf{a}\,dt + \sqrt{2\sigma_v^2}\,dW_v, \quad d\mathbf{a} = -\gamma_a\mathbf{a}\,dt + \sqrt{2\sigma_a^2\gamma_a}\,dW_a \tag{5a}$$

here W_v, W_a are independent Wiener processes. The discrete form (Euler-Maruyama, $\Delta t = 1$) is given in Eq. 5b.

$$\begin{aligned} \mathbf{v}_t &= \mathbf{v}_{t-1} + \mathbf{a}_{t-1} + \boldsymbol{\eta}_v, \quad \boldsymbol{\eta}_v \sim \mathcal{N}(0, \sigma_v^2\mathbf{I}), \\ \mathbf{a}_t &= \beta_a\mathbf{a}_{t-1} + \boldsymbol{\eta}_a, \quad \boldsymbol{\eta}_a \sim \mathcal{N}(0, \sigma_a^2\mathbf{I}), \\ \mathbf{x}_t &= \mathbf{x}_{t-1} + \mathbf{v}_t + \frac{1}{2}\mathbf{a}_{t-1}, \end{aligned} \tag{5b}$$

with $\beta_a = e^{-\gamma_a} = 0.741$, $\gamma_a = 0.3$(acceleration reversion rate), $\sigma_v = 0.12$(velocity noise scale), $\sigma_a = 0.03$(acceleration noise scale). Specifically, the acceleration state $a(t)$ evolves as a drift-controlled Ornstein-Uhlenbeck process [26]. Because the restoring force coefficient $\gamma_a > 0$ restricts unbounded variance expansion, the velocity variance over a finite path horizon H remains strictly bounded, preventing kinematic divergence.

Path Integral Formulation. The predictive posterior over future states (horizon H=5) is the Feynman-Kac solution to the controlled diffusion is given as follows.

$$p(\mathbf{x}_{t:H} \mid \mathbf{x}_{t-1}) \propto \exp \int \left(-\frac{1}{\tau} S[\xi]\right) \mathcal{D}\xi$$

where $\boldsymbol{\xi}$ enumerates paths from current state to $\mathbf{x}_{t:H}$, $\tau = 10.0$ is the temperature (controlling exploration vs. exploitation), and $S[\boldsymbol{\xi}]$is the composite action encoding priors. This formulation is derived from the Hamilton-Jacobi-Bellman equation under quadratic costs, weights paths by their physical and perceptual plausibility.

Composite Action Functional. The Action is a simple addition of various quantities given in Eq. 6.

$$S[\boldsymbol{\xi}] = S_{\text{kin}} + S_{\text{sem}} + S_{\text{roi}} + S_{\text{turn}} \tag{6}$$

Where, $S_{\text{kin}} = \sum_{h=1}^{H} \left(\| \mathbf{v}_{t+h} \|^2 + \frac{1}{2} \| \mathbf{a}_{t+h} \|^2 \right) \Delta t$ is the Kinetic/Control Cost (minimum energy principle) penalizing high speed and jerk to favor inertial motion, $S_{\text{sem}} = \lambda_{\text{sem}}(1 - \mathbf{e}(\mathbf{x}_{t+H})^{\top} \mathbf{e}_{\text{ref}})$ is the Semantic/Appearance Cost where $\mathbf{e}(\mathbf{x}_{t+H})$is CLIP embedding of a 32×32 patch at the endpoint, and $\lambda_{\text{sem}} = 5.0$, $S_{\text{roi}} = \lambda_{\text{roi}} \sum_{h=1}^{H} \text{softplus}\big(m - d_{\text{roi}}(\mathbf{x}_{t+h})\big)\Delta t$ is the Region of Interest (ROI) Boundary avoidance Cost is a soft physical constraint that heavily penalizes any predicted future trajectory that would leave the visible image frame with $\lambda_{\text{roi}} = 120m$ m=5px (margin) and $d_{\text{roi}}(\mathbf{x})$ = min(dist to boundary edges) (> 0 inside ROI, < 0 outside), and $S_{\text{turn}} = \lambda_{\text{turn}} \sum_{h=1}^{H} \text{softplus}(\quad \mid \angle(\mathbf{v}_{t+h-1}, \mathbf{v}_{t+h}) \mid -\theta_{\max})\Delta t$ is the turn penalty which is a soft physical prior that strongly discourages unrealistically sharp direction changes like instant 90° turns and $\lambda_{\text{turn}} = 12.0$, $\theta_{\max} = \frac{\pi}{4}$, and $\angle(\mathbf{u}, \mathbf{v}) = \arccos\left(\frac{\mathbf{u}^{\top}\mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}\right)$.

Monte Carlo Approximation and Prediction. We approximate Feynman-Kac solution with N = 30 forward sampled trajectories (Eq. 5b.), computing weights for each track by Eq. 7.

$$w_k = \exp\left(-\frac{S_k - S_{\min}}{\tau}\right) \Big/ {}_k \text{, where } k = \sum\nolimits_j \exp\left(-\frac{S_j - S_{\min}}{\tau}\right) \tag{7}$$

Where, S_k is the total Action cost of the k-th sampled future path, S_{min} is the Action of the best (lowest cost) of the N sampled paths, the importance weight of each sampled trajectory follows the classic Gibbs/Boltzmann form (Eq. 7). By subtracting $S_{\min}$we avoid numerical overflow, and the temperature $\tau = 10.0$ yields a soft decisive selection where physically implausible paths (high Action score) like high turn penalty, boundary violation, or large kinetic cost are exponentially down-weighted, while smooth (low Action score), in-frame trajectories dominate the predictive distribution. The mapping of $S[\xi]$ to $Eq.7.$ satisfies the localized Feynman-Kac transformation [11]. This transformation guarantees that the trajectory distribution smoothly linearizes the underlying non-linear Hamilton-Jacobi-Bellman (HJB) equations, formally ensuring that the state estimation posterior remains structurally stable and converges on the path of least action.

The hyperparameter values are locked based on foundational Monte Carlo numerical simulations evaluating motion under measurement noise and synthetic occlusions.

The predicted position is simply a weighted sum, shown in Eq. 8.

$$\hat{\mathbf{x}}_{t|t-1} = \sum_{k=1}^{N} w_k \mathbf{x}_{t+H}^{(k)} \tag{8}$$

Measurement Update. Upon successful association, we perform a physics-consistent measurement update, the position $\mathbf{x}_t$ is immediately corrected to the observed detection $\mathbf{z}_t$ (hard update), while velocity is only softly adjusted using a momentum-preserving moving average $\mathbf{v}_t \leftarrow 0.8\mathbf{v}_{t-1} + 0.2\frac{\mathbf{z}_t - \mathbf{x}_{t-1}}{\Delta t}$. This respects physical inertia (no instantaneous velocity jumps) and is the direct discrete-time analogue of the "control" step in stochastic optimal control. By resetting the positional state to the high-confidence detection while smoothing the first-order derivative (velocity), PIPF eliminates the cumulative drift inherent in recursive filters without sacrificing the smoothness of the trajectory. In Fig. 5 shows that our PIPF motion model beats the standard motion model of Kalman at a sharp 90° turn, of GT, made synthetically.

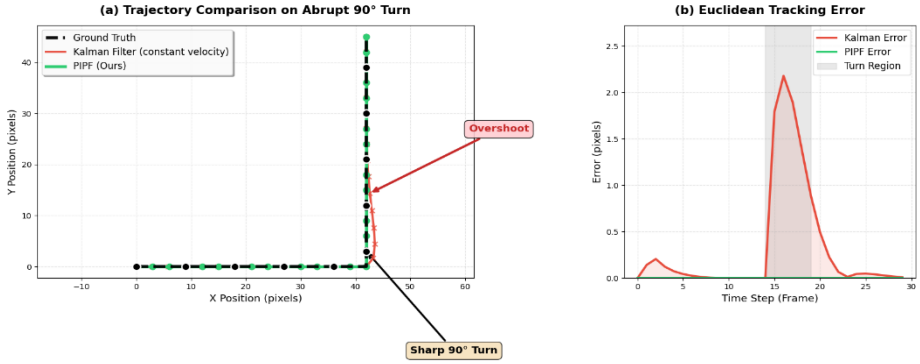

Fig. 5. (a) This indicates how Kalman overshoots at a sharp turn while PIPF does not. (b) shows PIPF gives zero error though out, Kalman error explodes at the turn. (Use 200% zoom for better visibility)

3.4 Global Data Association and Track Management

At frame t, let $\mathcal{T}_{t-1}$ be active tracks and $\mathcal{D}_t = \{\mathbf{z}_t^j, \mathbf{e}_t^j\}_{j=1}^{N_t}$ be detections. We form a cost matrix $C \in \mathbb{R}^{|\mathcal{T}_{t-1}| \times N_t}$ as Eq. 9.

$$C_{ij} = \left(1 - \mathbf{e}_i^{\top}\mathbf{e}_j\right) + 0.5 \cdot \frac{\| \hat{\mathbf{x}}_i - \mathbf{z}_t^j \|_2}{\boldsymbol{\sigma}_{diag}} \tag{9}$$

where σ_{diag} represents the diagonal image, rendering the metric invariant to video resolution, its value is taken to be 1000 for steady computation purposes. Optimal assignment is done via Hungarian Algorithm (Jonker-Volgenant), shown in Eq. 10.

$$\pi^* = \arg\min_{\pi} \sum_{i} C_{i,\pi(i)} \tag{10}$$

Matched tracks are updated, unmatched tracks coast via PIPF, gain age every time, until age = 45, and are terminated.

For better understanding, pseudocode of the proposed method is given in Algorithm 1.

Algorithm 1: Physics-Informed Path Integral Filter (PIPF)

Input: Video sequence $\mathcal{V} = \{I_t\}_{t=1}^{T}$; Frozen Detector D; Frozen Encoder Φ; Hyperparameters: τ, T, N, α_e.
Output: Set of trajectories $\mathcal{T}$.

1. **Initialization:** Set active tracks $\mathcal{T}_{active} \leftarrow \emptyset, globaltrackID\ k \leftarrow 0$.
2. **For each frame** $I_t \in \mathcal{V}$ **do**:
   - **Detection & Feature Extraction:**
     - Extract polygons $\{b_t^i\}$ using D. Filter detections by confidence score
     - For each polygon b_t^i, compute normalized visual embedding e_t^i via Φ.
   - **Physics-Informed Motion Forecasting (PIPF):**
     - **For each track** $\tau_k \in \mathcal{T}_{active}$ **do:**
       - Initialize N particles from current state $(x_{t-1}, v_{t-1}, a_{t-1})$
       - **For each particle** $n \in \{1 \dots N\}$ **do:**
         - Sample future paths via discretized SDE over horizon T
         - Calculate Composite Action $\boldsymbol{S}[\boldsymbol{\xi_n}]$
       - Compute importance weights $\boldsymbol{w_n}$ as eq. 7
       - Forecast predicted position $\hat{\mathbf{x}}_{t|t-1}$
     - **Global Data Association:**
       - Construct Cost Matrix $\boldsymbol{C}$
       - Solve optimal assignment $\boldsymbol{\pi}^*$ via the Hungarian Algorithm
     - **Track Management & State Update:**
       - **Matched Tracks:**
         - Update position $x_t \leftarrow z_{\pi^*}$ (hard update).
         - Soft-update velocity $v_t \leftarrow 0.8v_{t-1} + 0.2\frac{z_t - x_{t-1}}{\Delta t}$ to respect inertia.
     - **Unmatched Detections:** Initialize new tracks, age = 0; $k \leftarrow k + 1$.
     - **Unmatched Tracks:** Increment track age; "coast" using the PIPF prediction.
     - **Termination:** Delete tracks where age > 45 frames.
3. **Return Final trajectories** $\mathcal{T}$

4 Experimental results

4.1 Datasets and Evaluation Metrices

We evaluate two standard video text benchmarks: **ICDAR 2015 Text in Video,** which contains 49 videos (25 Training and 24 Test set videos), and contains relatively smooth camera motion, sparse text, etc., **ICDAR 2023 DSText,** which contains 100 videos (50 Training and 50 Test). The second dataset is relatively more challenging and contains dense and small text, like gaming, sports videos, etc. It also often shows heavy motion and blur, making it an ideal test data for zero-shot methods.

Performance is measured using the official CLEAR MOT metrics, which are MOTA, MOTP, and IDF1 [24, 25]. All metrics are computed on the official evaluation servers using submitted XML tracks. To quantify the trade-off between performance and annotation cost, we introduce the Video Supervision Efficiency (VSE) index. VSE

penalises raw accuracy based on the logarithmic volume of video-specific training data required, modelling the diminishing returns of large-scale supervision as defined in Equation (11).

$$VSE = \frac{\text{MOTA}}{\log_{10}(10 + N_{frames})} \tag{11}$$

where N_{frames} is the number of annotated frames in the video training set, its log parametrization in the denominator is directly grounded in sample complexity bounds [27] and empirical risk scaling laws [28]. This metric isolates the marginal value of video supervision, where a higher VSE indicates a more data-efficient tracking paradigm. For ICDAR 2015 Video $N_{frames} \approx 13{,}400$, ICDAR 2023 DSText $N_{frames} \approx 28{,}000$. As PIPF requires no video training, therefore $N_{frames} = 0$. All experiments are conducted on a single RTX 2080 Ti (11 GB). No video-specific training is performed by us at any stage.

4.2 Plug-and-Play Experimentation results

To isolate the novelty and effectiveness of PIPF, and to demonstrate its generalization capability, we conducted experiments by swapping detector models and embedding models.

Table 1. Results of the plug-and-play experimentation.

Experiments	ICDAR 2015			ICDAR 2023 DSText		
Measures	MOTA	MOTP	IDF1	MOTA	MOTP	IDF1
DBnet + CLIP + PIPF	24.11	73.41	43.64	23.82	75.51	41.93
CRAFT + DINO + PIPF	**29.05**	**73.73**	**49.03**	21.74	71.02	41.55
DBnet + DINO + PIPF	25.86	73.41	47.18	**26.39**	**75.51**	**46.73**
CRAFT + CLIP + PIPF (Baseline)	27.45	73.72	43.63	20.16	71.03	37.75

To validate PIPF as a model-agnostic motion interface, we systematically swapped upstream detectors (CRAFT [9] vs. DBNet++ [10]) and semantic embedders (CLIP [7] vs. DINOv2 [13]). This plug-and-play evaluation isolates PIPF's role, confirming it translates static model upgrades directly into tracking gains without modifying the path integral motion model, as shown in Table 1.

Superiority of Self-Supervised Features: Across benchmarks, DINOv2 consistently outperforms CLIP on ICDAR 15 and DSText in terms of IDF1. This highlights a key insight for zero-shot tracking: DINOv2's patch-level, geometrically self-supervised features provide finer-grained discrimination for visually similar text under occlusion, surpassing CLIP's global, language-aligned vectors. **Detector Sensitivity and Domain Fit**: PIPF's performance is bounded by the detector's domain alignment, but the motion model adapts robustly. On ICDAR 2015, CRAFT's character-level affinities excel for curved text compared to DBNet++ in terms of MOTA. On dense DSText, DBNet++ prevails over CRAFT in terms of MOTA. This shows PIPF as a stable interface that scales with detector quality via simple frontend swaps. **New Zero-Shot Baseline**: The DBNet++ + DINOv2 combo sets a zero-shot SOTA, with 25.86 MOTA / 47.18 IDF1

on ICDAR 2015 and 26.39 MOTA / 46.73 IDF1 on DSText vs. baselines like CRAFT + CLIP (20.16 MOTA / 37.75 IDF1). This establishes PIPF as a scalable Data Engine, generating high-fidelity silver-standard trajectories from raw videos at zero training cost.

4.3 Comparative study

We compared PIPF with some SOTA, shown separately for the two benchmarks.

Table 2. Comparison on ICDAR 2015 video data

Methods	Supervision	IDF1 ↑	MOTP ↑	MOTA ↑	VSE↑
AJOU$^{\Delta}$	Video-trained	36.07	72.71	16.44	3.98
SRCB$^{\Delta}$	Video-trained	39.4	68.51	23.09	5.59
$HIK_{OCR}{}^{\Delta}$	Video-trained	57.92	76.78	43.16	10.45
TransDETR$^{\Delta}$ [2]	Video-trained	65.33	74.01	48.86	11.83
CoText*[3]	Video-trained	67.4	73.6	50.3	12.18
GoMatching++ [4]	Video-trained	**72.87**	**77.94**	**62.41**	15.12
CRAFT + DINO + PIPF	**No training**	49.03	73.73	29.05	**29.05**
CRAFT + CLIP + PIPF (Baseline)	**No training**	43.63	73.72	27.45	27.45

Table 3. Comparison on DSText data

Methods	Supervision	IDF1 ↑	MOTP ↑	MOTA ↑	VSE↑
TextTrack$^{\Delta}$	Video-trained	50.22	74.03	25.75	5.79
TransDETR* [2]	Video-trained	55.61	77.15	39.98	8.99
GoMatching++[4]	Video-trained	58.44	**80.01**	48.0	10.79
TencentOCR$^{\Delta}$	Video-trained	**75.87**	79.88	**62.56**	14.07
DBnet + DINO + PIPF	**No training**	46.73	75.51	26.39	**26.39**
CRAFT + CLIP + PIPF (Baseline)	**No training**	37.75	71.03	20.16	20.16

We compare PIPF against supervised SOTA methods on ICDAR 2015 Video and 2023 DSText benchmarks in Tables 2 and 3, respectively. All baselines are trained end-to-end on video splits (~13k frames for ICDAR15, ~28k for DSText), while PIPF uses zero video training. The PIPF achieves the best VSE on both datasets compared to existing models. This justifies that the proposed model is efficient due to zero video training. However, the PIPF closes a sufficient IDF1 gap to SOTA, performing better than some old supervised methods, demonstrating physics priors' power for zero-shot pattern recognition in motion. PIPF achieves almost equal MOTP, falls short of outperforming the SOTA on MOTA and IDF1 scores because it is completely untrained on the specific video sequence domain, but still achieves comparable results to the SOTA, whereas in Data efficiency, PIPF performs almost double better to that of the supervised models. Therefore, one can conclude that the proposed model is very data-efficient but at the cost of accuracy. Also, PIPF is Detector-Agnostic, a better off the shelf can be plugged into it to improve the results, as shown in Table 1, leaving room for improvement in future. ($Scores\ taken:$ $^{\Delta}$ICDAR leaderboards , *from [4])

4.4 Ablation Study

To validate the contributions, we conducted an ablation study using two zero-shot variants with our PIPF on the same CRAFT detection and CLIP embeddings pipeline. The variants are designed to be orthogonal in purpose: (i) **Kalman Motion.** Uses pure geometric motion modelling via a constant-velocity Kalman filter with IoU-based Hungarian association, (ii) **Appearance-Hungarian.** It uses semantic discrimination using CLIP cosine similarity with Hungarian assignment and exponential embedding smoothing. It is essentially PIPF without its motion model, (iii) **PIPF (Ours).** It integrates both orthogonal aspects via physics-informed path integral control (Section 3.3), where motion priors guide predictions, and CLIP embeddings inform association. Results are reported separately for each dataset to highlight domain-specific behaviours (Table 4).

Table 4. Ablation on the ICDAR 2015 video and DSText

Steps	ICDAR 2015				ICDAR 2023 DSText			
Measures	MOTA↑	MOTP↑	IDF1↑	FPS↑	MOTA↑	MOTP↑	IDF1↑	FPS↑
Kalman Motion	-10.86	73.61	36.64	3.96	-31.6	70.25	17.66	0.25
Appearance-Hungarian	22.65	73.67	40.8	**6.1**	15.87	70.96	35.02	**3.59**
Proposed	**27.45**	**73.72**	**43.63**	0.49	**20.16**	**71.03**	**37.75**	1.76

On both benchmarks, the constant-velocity Kalman baseline collapses to −10.86 / −31.6 MOTA, indicating thousands of identity switches because linear extrapolation cannot cope with handheld jerk, sharp turns, or border cropping. Replacing motion with pure CLIP-based appearance + Hungarian assignment recovers by exploiting CLIP's remarkable invariance to blur, scale, and partial occlusion. However, without any notion of inertia or velocity, it still suffers severe trajectory fragmentation in long occlusions and crowded scenes, yielding 15–22 % MOTA and IDF1 limited to ~35–40, performing better than the Kalman.

PIPF outperforms both ablations by +4.8 MOTA / +4.3 MOTA and +2.8 / +2.7 IDF1 points than the 2nd best indicates the least identity switches of the three as CRAFT and CLIP (when applicable) remains the same. These gains arise directly from the path integral's exponential weighting (Eq. 7), which concentrates probability mass on low-action trajectories that respect kinetic smoothness, bounded turning and image boundary avoidance. Also, on DSText benchmark for increasing FPS and disabling forward CLIP prediction still yields +4.3 MOTA over appearance-only, proving that pure physical priors are sufficient for robust dynamics modelling. This better performance comes at the cost of FPS, as shown in Table 4. All the ablations use the same CRAFT model, which is not fine-tuned on the evaluation data. Due to this zero-shot nature, overall, MOTA remains low, leaving room for improvement.

4.5 Limitations

Although PIPF is fully zero-shot and already beats few supervised trackers, two practical aspects remain open: both are engineering opportunities rather than fundamental barriers: (i) Inference speed in ultra-dense scenes DSText frequently exceeds 100 simultaneous text instances. Our Monte Carlo sampling (N=30 paths per active

track) currently yields 1.76 FPS on an RTX 2080 Ti. Remarkably, activating the adaptive 'Pure Physics' mode (λ_{sem} =0) not only stabilizes association in extreme clutter but doubles speed. This confirms that our physical priors alone are sufficient in the hardest regimes, paving the way for lightweight distillation of the path-integral policy into a <5 ms neural proxy for 15 to 25 FPS real-time deployment (ii) Upper bound by single-frame detection like all tracking-by-detection paradigms, PIPF inherits the recall of the frozen CRAFT detector. Prolonged severe motion blur or low-contrast text still causes missed detections and fragmentation. However, because our motion model is training-free and physically grounded, it naturally invites plug-and-play upgrades with any future zero-shot or weakly supervised text spotter without retraining the tracker itself.

5 Conclusion and Future Work

We presented PIPF, a training-free Physics-Informed Path Integral Filter that achieves competitive video text tracking solely from frozen static detectors and foundation-model embeddings. By formulating motion forecasting as stochastic optimal control under classical mechanics priors, PIPF achieves comparable results with double data efficiency over the supervised models. Our results validate an underexplored yet powerful direction: first-principles physics combined with modern foundation models can serve as an interpretable, instantly deployable, and infinitely upgradable alternative to massive end-to-end video supervision. PIPF not only closes a large portion of the performance gap but also acts as a scalable Data Engine that turns any unlabelled video archive into high-quality trajectory annotations for free. Future work includes distilling PIPF for real-time deployment, into a lightweight, feed-forward neural proxy network (e.g., a fast MLP or small RNN policy). We believe physics-informed paradigms will play a central role in the next generation of efficient and trustworthy video perception systems.

References

1. Cheng, Z., Lu, J., Zou, B., Qiao, L., Xu, Y., Pu, S., Niu, Y., Wu, F., Zhou, S.: FREE: A Fast and Robust End-to-End Video Text Spotter. IEEE TIP 30, 822–837 (2021)
2. Wu, W., Cai, Y., Shen, C., Zhang, D., Fu, Y., Zhou, H., Luo, P.: End-to-End Video Text Spotting with Transformer. arXiv preprint arXiv:2203.10539 (2022)
3. Wu, W., Li, Z., Li, J., Shen, C., Zhou, H., Li, S., Wang, Z., Luo, P.: Real-time End-to-End Video Text Spotter with Contrastive Representation Learning. arXiv preprint arXiv:2207.08417 (2022)
4. He, H., Zhang, J., Ye, M., Liu, J., Du, B., Tao, D.: GoMatching++: Parameter- and Data-Efficient Arbitrary-Shaped Video Text Spotting and Benchmarking. arXiv preprint arXiv:2505.22228 (2025)
5. Wu, W., Zhao, Y., Li, Z., Li, J., Shou, M.Z., Pal, U., Karatzas, D., Bai, X.: ICDAR 2023 Video Text Reading Competition for Dense and Small Text. arXiv preprint arXiv:2304.04376 (2023)

6. Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V.R., Lu, S., et al.: ICDAR 2015 Competition on Robust Reading. pp. 1156–1160. IEEE (2015)
7. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I, In Proc. ICML pp. 8748–8763. (2021)
8. Yuan, L., Chen, D., Chen, Y.L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al.: Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432 (2021)
9. Baek, Y., Lee, B., Han, D., Yun, S., Lee, H.: Character Region Awareness for Text Detection. In Proc. CVPR pp. 9365–9374 (2019)
10. Liao, M., Zou, Z., Wan, Z., Yao, C., Bai, X.: Real-Time Scene Text Detection with Differentiable Binarization and Adaptive Scale Fusion. IEEE TPAMI 919–931 (2023)
11. Kappen, H.J.: Linear theory for control of nonlinear stochastic systems. Physical Review Letters 95(20), 200201 (2005)
12. Theodorou, E., Buchli, J., Schaal, S.: A generalized path integral control approach to reinforcement learning. Journal of Machine Learning Research **11**, 3137–3181 (2010)
13. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning Robust Visual Features without Supervision. Transactions on Machine Learning Research (TMLR) (2024)
14. Liao, M., Shi, B., Bai, X.: TextBoxes++: A Single-Shot Oriented Scene Text Detector. IEEE Transactions on Image Processing 27(8), 3676–3690 (2018)
15. Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., Liang, J.: EAST: An Efficient and Accurate Scene Text Detector. In Proc. CVPR pp. 2642–2651. (2017)
16. Li, X., Wang, W., Hou, W., Liu, R.-Z., Lu, T., Yang, J.: Shape Robust Text Detection with Progressive Scale Expansion Network. In Proc. CVPR pp. 9336–9345 (2019)
17. Liao, M., Wan, Z., Yao, C., Chen, K., Bai, X.: Real-time Scene Text Detection with Differentiable Binarization. In Proc. AAAI pp. 11474–11481 (2020)
18. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In Proc. ICIP pp. 3464–3468. (2016)
19. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In Proc. ICIP pp. 3645–3649. (2017)
20. Ratliff, N., Zucker, M., Bagnell, J.A., Srinivasa, S.: CHOMP: Gradient optimization techniques for efficient motion planning. In Proc. ICRA pp. 489–494. (2009)
21. Kamtue, K., Moura, J.M., Sangpetch, O., Garcia, P.: PhyOT: Physics-informed object tracking in surveillance cameras. In Proc. ICASSP pp. 1–5. (2024)
22. Gupta, A., Vedaldi, A., Zisserman, A.: Synthetic Data for Text Localisation in Natural Images. In Proc. CVPR, pp. 2315–2324 (2016)
23. Nayef, N., Yin, F., Bizid, I., Choi, H., Feng, Y., Karatzas, D., Luo, Z., Pal, U., Rigaud, C., Chazalon, J., et al.: ICDAR2017 Robust Reading vol. 1, pp. 1454–1459. (2017)
24. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: The CLEAR MOT metrics. EURASIP Jour. on Image and Video Processing 2008, 1–10 (2008)
25. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In Proc. ECCV. pp. 17–35. Springer (2016)
26. Gardiner, C.: Stochastic Methods: A Handbook for the Natural and Social Sciences. 4th edn. Springer, Berlin (2009)
27. Vapnik, V.N.: Statistical Learning Theory. John Wiley & Sons, New York (1998)
28. Kaplan, J., et al.: Scaling laws for neural language models. arXiv preprint arXiv:2001.08361 (2020)