

Visual-Inertial Motorcycle State Estimation Using a BiLSTM-Transformer Architecture

Mohamed Sanim Akremi¹ Hicham Hadj-Abdelkader² Lamri Nehaoua³ Hichem Arioui⁴

Abstract—An accurate real-time estimate of motorcycle dynamic states is essential for advanced riding assistance systems (ARAS) aimed at reducing the risk of the rider. Traditional model-based observers require precise vehicle parameters and struggle with the inherent nonlinearities of motorcycle behavior, particularly during low-speed steady turns and aggressive transients. This paper presents an enhanced deep learning framework that fuses monocular visual data with inertial measurements to estimate key dynamic states, including steering angle, lateral velocity, lateral offset, and tire slip angles. Building upon a baseline ResNet50-LSTM architecture, we propose a BiLSTM-Transformer hybrid that captures richer temporal dependencies through bidirectional processing and multi-head self-attention. The model processes sequences of images and inertial signals without requiring prior knowledge of the vehicle model. Experimental validation on synthetic data demonstrates that the proposed architecture significantly reduces estimation errors, particularly when addressing the steady-state offset observed at low speeds. The method maintains real-time performance while providing attention-based interpretability, representing a step toward more reliable and deployable ARAS for motorcycles.

Index Terms—Motorcycle dynamics, BiLSTM, Transformer, state estimation, ARAS, visual-inertial fusion.

I. INTRODUCTION

Powered two-wheelers remain vulnerable road users because stability, braking, steering, and rider body motion are strongly coupled. Modern motorcycles increasingly include cornering ABS, traction control, and radar- or camera-based assistance functions, but these systems require reliable knowledge of the motorcycle state to warn or intervene safely. Unlike cars, motorcycles exhibit strongly coupled dynamics in which lean angle, steering angle, and lateral velocity are closely linked [2]. Effective riding also requires simultaneous management of balance, steering, braking, and road perception, making accurate assistance particularly important [11].

Advanced rider assistance systems increasingly address cornering stability, traction management, obstacle detection, and anticipatory interventions. Recent industrial developments combine inertial, radar, and camera sensing for functions such as collision warning, blind-spot detection, lane detection, and lane departure warning [14]. These developments

reinforce the need for robust real-time estimation algorithms able to provide accurate motorcycle dynamic states.

A fundamental challenge is that several key states, particularly tire slip angles and lateral velocity, are not directly measurable with standard low-cost sensors. Front and rear slip angles are especially relevant for characterizing lateral stability, yet they depend on tire-road friction, tire condition, inflation pressure, and vehicle loading. Traditional model-based observers address this problem using, for example, unknown-input observers for Linear Parameter-Varying systems [3], [8] and Takagi-Sugeno approaches for lean and steering dynamics estimation [8]. However, these methods require sufficiently accurate knowledge of vehicle parameters such as mass distribution, suspension geometry, and tire characteristics, which may vary in practice.

Recent work has therefore explored machine learning for vehicle and motorcycle state estimation. Long Short-Term Memory (LSTM) networks have been used to predict motorcycle trajectory from dynamic states, rider inputs, and body movement [9]. Transformer-based approaches have also shown strong potential for vehicle dynamics modeling by exploiting self-attention to capture temporal dependencies [10]. In parallel, vision-based methods have demonstrated that monocular cameras can provide useful information for motorcycle attitude, lateral position, steering behavior, and road geometry estimation [1], [4]. These developments motivate combining visual information with inertial measurements in a unified temporal learning framework.

The present work builds upon previous visual-inertial motorcycle estimation studies by introducing a BiLSTM-Transformer architecture that combines ResNet50 image features with inertial measurements over an eight-frame temporal window. The main contributions of this paper are as follows:

- 1) A bidirectional Long Short-Term Memory and Transformer hybrid architecture for motorcycle lateral dynamics estimation that captures richer temporal dependencies than unidirectional recurrent networks.
- 2) Multi-modal fusion of visual and inertial data using low-cost sensors without requiring prior vehicle model knowledge.
- 3) Attention-based interpretability, enabling visualization of which temporal frames most influence each state prediction.
- 4) Real-time performance maintaining standard video frame rate inference capability suitable for advanced rider assistance systems deployment.

¹Mohamed Sanim Akremi is a postdoc at IBISC laboratory, University of Evry Paris-Saclay mohamed.akremi@universite-paris-saclay.fr

²Hicham Hadj-Abdelkader is an associate professor at University of Evry Paris-Saclay hicham.hadjabdelkader@univ-evry.fr

³Lamri Nehaoua is an associate professor at University of Evry Paris-Saclay lamri.nehaoua@univ-evry.fr

⁴Hichem Arioui is a full Professor at University of Evry Paris-Saclay hichem.arioui@univ-evry.fr

II. RELATED WORK

A. Artificial Intelligence for Vehicle State Estimation

Long Short-Term Memory networks, a specialized form of recurrent neural networks, have proven particularly effective for temporal state estimation tasks where current states depend on past observations. For trajectory prediction in motorcycles, researchers have employed parallel LSTM layers processing both high-dynamic and low-dynamic input signals including motorcycle internal measurements, rider inputs, and body and head movement to predict roll angle and velocity over prediction horizons of four seconds [9]. This work demonstrated that LSTM-based prediction of lateral dynamics achieves significant improvements over constant-curvature baseline methods, with up to thirty-seven percent improvement in longitudinal root mean square error when combining roll angle and velocity prediction [9].

More recently, Transformer architectures, originally developed for natural language processing, have been adapted to temporal modeling tasks in vehicle dynamics. The self-attention mechanism of Transformers enables the network to selectively focus on the most informative time steps, potentially capturing long-range dependencies that recurrent networks may miss. A notable advance is the development of residual learning approaches for vehicle dynamics modeling using Transformers, where a Transformer-based dynamics residual correction network, called DyTR, is designed to correct the state residuals of a physical model rather than directly estimating the states [10]. This residual learning strategy greatly reduces the difficulty of network learning and improves estimation accuracy. Experiments demonstrated that this approach reduced state prediction errors of a simple three-degree-of-freedom vehicle model by an average of 92% and 60% across two datasets, respectively. This result suggests that combining physical models with Transformer-based corrections offers a promising direction for high-fidelity vehicle state estimation.

The combination of convolutional and recurrent architectures has also emerged as a promising direction for vision-based vehicle state estimation. Convolutional neural networks can extract spatial features from images while recurrent networks capture temporal evolution, enabling robust estimation even in challenging conditions. Recent doctoral research at the IBISC laboratory has specifically explored vision-based estimation for motorcycles, developing an Inverse Perspective Mapping method for estimating full vehicle attitude relative to the road, as well as a more efficient Direct Pixel Mapping technique operating directly on image pixels [1]. Furthermore, geometric approaches were developed to estimate lateral dynamics, including side-slip angle and lateral velocity, by analyzing vanishing points and lane geometry [1]. These vision-based algorithms, validated through extensive simulation, demonstrated high accuracy with sub-degree attitude estimation and real-time viability on commodity hardware [1].

B. Motorcycle Assistance Systems and State Estimation

Research on motorcycle assistance systems follows two main trajectories: model-based observer design and data-driven approaches. Traditional model-based methods, developed largely at the IBISC laboratory, include unknown input observers that estimate lateral dynamics independently of unknown inputs like rider torque, using Linear Matrix Inequality constraints for robust estimation [3], [8]. Takagi-Sugeno and Linear Parameter-Varying approaches have also been applied to lean and steering dynamics [8], [3]. However, these methods require precise knowledge of vehicle parameters (mass, inertia, tire characteristics) that are often unavailable.

In parallel, machine learning has been explored for motorcycle estimation. LSTM networks have successfully predicted trajectory using internal signals and rider movement data from 74 hours of open-road riding by 21 riders [9]. Vision-based systems have examined rider gaze behavior, confirming higher cognitive load and distinct gaze strategies for motorcyclists [1]. Integrating vision with inertial sensors, recent work shows a low-cost monocular camera can estimate attitude and lateral dynamics without vehicle-specific parameters [1]. The present work builds on this by proposing a BiLSTM-Transformer fusing visual and inertial data for real-time estimation without prior vehicle knowledge.

III. PROPOSED METHODOLOGY

This section presents the proposed deep learning framework for estimating motorcycle dynamic states from visual and inertial data. The core idea is to combine convolutional neural networks for spatial feature extraction, bidirectional recurrent processing for temporal context, and self-attention mechanisms for learning which moments in the sequence matter most for each output state. The overall architecture processes sequences of eight consecutive frames and outputs estimates of steering angle, lateral velocity, lateral offset, and front and rear tire slip angles.

A. Problem Formulation

Let us consider a motorcycle equipped with a monocular camera and a suite of low-cost inertial sensors. The camera is mounted on the headlight assembly and inclined downward by a tilt angle μ to capture the road ahead. The inertial suite comprises a six-axis inertial measurement unit (providing three-axis accelerometer and three-axis gyroscope measurements), a magnetometer, and a Global Navigation Satellite System receiver. All inertial measurements are taken from the rear chassis mass center, which serves as a stable reference point.

At each discrete time step t , we acquire the following measurements:

- An RGB image $\mathbf{I}_t \in \mathbb{R}^{H \times W \times 3}$, (fixing here as example $H = 224$ and $W = 224$)
- Longitudinal acceleration $a_x(t) \in \mathbb{R}$
- Lateral acceleration $a_y(t) \in \mathbb{R}$
- Angular velocities $\omega_x(t), \omega_y(t), \omega_z(t) \in \mathbb{R}$

- Longitudinal velocity $v_x(t) \in \mathbb{R}$ provided by the Global Navigation Satellite System

We define the input vector $\mathbf{s}$ at time t as:

$$\mathbf{s}_t = [a_x(t), a_y(t), \omega_x(t), \omega_y(t), \omega_z(t), v_x(t)]^\top \in \mathbb{R}^6 \quad (1)$$

The objective is to estimate the following five dynamic states at each time step:

$$\mathbf{y}(t) = [\delta(t), v_y(t), \Delta_y(t), \alpha_f(t), \alpha_r(t)]^\top \in \mathbb{R}^5 \quad (2)$$

where $\delta(t)$ is the steering angle, $v_y(t)$ is the lateral velocity, $\Delta_y(t)$ is the lateral offset relative to the lane center, and $\alpha_f(t)$ and $\alpha_r(t)$ are the front and rear tire slip angles, respectively [2].

Unlike traditional model-based approaches that require explicit knowledge of vehicle parameters [3], [8], we seek to learn a mapping function f_θ parameterized by a neural network with weights θ :

$$\hat{\mathbf{y}}(t) = f_\theta(\mathbf{I}_{t-K+1:t}, \mathbf{s}_{t-K+1:t}) \quad (3)$$

where $K = 8$ is the length of the temporal window, and $\mathbf{I}_{t-K+1:t}$ and $\mathbf{s}_{t-K+1:t}$ denote sequences of K consecutive images and inertial measurements. The mapping is learned directly from data without assuming any specific vehicle model.

B. Overview of the Proposed Architecture

Figure 1 presents a comprehensive overview of the proposed BiLSTM-Transformer architecture. The network processes each eight-frame sequence through four main stages. First, a ResNet50 convolutional neural network extracts visual features from each image independently. Second, the visual features are concatenated with the corresponding inertial measurements for each frame. Third, a bidirectional Long Short-Term Memory layer processes the sequence, capturing both past and future context [7]. Fourth, a Transformer encoder with multi-head self-attention learns which time steps are most relevant for each prediction [15]. Finally, global average pooling aggregates information across the sequence, and fully connected layers regress the five output states.

The design builds upon recent advances in visual-inertial state estimation for motorcycles [13], [1], [4], extending the temporal modeling capacity through bidirectional processing and self-attention.

C. Visual Feature Extraction with ResNet50

For each image $\mathbf{I}_t$ in the input sequence, we employ a ResNet50 network pre-trained on the ImageNet database [6]. ResNet50 was selected after evaluating several candidate architectures including ResNet18, ResNet101, and VGG variants for its optimal balance between representational power and training efficiency. The network consists of 50 layers organized into five stages, with skip connections that mitigate the vanishing gradient problem and enable effective training of deep architectures.

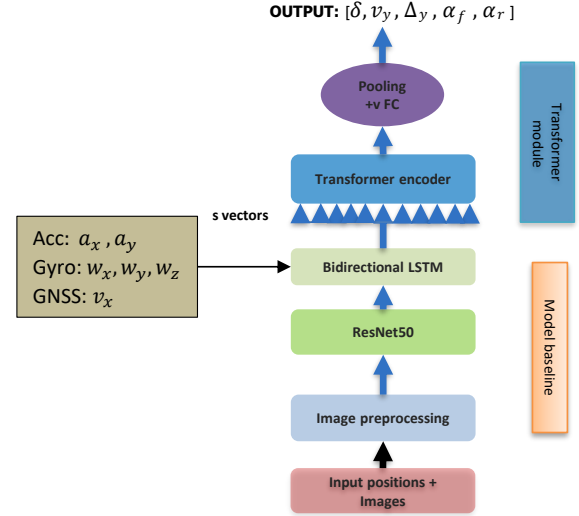


Fig. 1. Proposed BiLSTM-Transformer architecture for motorcycle lateral dynamics estimation. The network processes 8-frame sequences of images and inertial data to estimate five dynamic states.

Let ϕ_{ResNet} denote the ResNet50 backbone with the final classification layer removed. The visual feature extraction for frame t is:

$$\mathbf{f}_t = \phi_{\text{ResNet}}(\mathbf{I}_t) \in \mathbb{R}^{512} \quad (4)$$

The output is a 512-dimensional feature vector that encodes the visual appearance of the road scene, including lane markings, road curvature, and the position of the motorcycle relative to the lane.

D. Temporal Feature Fusion with Bidirectional LSTM

After extracting visual features, we concatenate them with the inertial measurements to form a combined feature vector for each time step:

$$\mathbf{x}_t = [\mathbf{f}_t; \mathbf{s}_t] \in \mathbb{R}^{518} \quad (5)$$

The sequence $\mathbf{x}_{1:8} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_8\}$ is then processed by a bidirectional Long Short-Term Memory network [7]. Unlike standard unidirectional LSTM that processes time steps from past to future only, the bidirectional variant maintains two separate recurrent passes:

$$\vec{\mathbf{h}}_t = \text{LSTM}_{\text{fw}}(\mathbf{x}_t, \vec{\mathbf{h}}_{t-1}) \quad (\text{forward pass}) \quad (6)$$

$$\overleftarrow{\mathbf{h}}_t = \text{LSTM}_{\text{bw}}(\mathbf{x}_t, \overleftarrow{\mathbf{h}}_{t+1}) \quad (\text{backward pass}) \quad (7)$$

Both LSTM passes have 128 hidden units. The forward pass encodes information from earlier frames, while the backward pass encodes information from later frames. At each time step, the two hidden states are concatenated:

$$\mathbf{H}_t = [\vec{\mathbf{h}}_t; \overleftarrow{\mathbf{h}}_t] \in \mathbb{R}^{256} \quad (8)$$

This bidirectional processing is particularly valuable for motorcycle state estimation because the upcoming road

geometry influences current steering decisions. As a rider approaches a curve, they begin steering before entering the curve, meaning that future frames contain information relevant to present states [1], [4]. The bidirectional architecture learns this anticipation effect directly from data without requiring explicit prediction of future road curvature.

E. Positional Encoding

The output of the bidirectional LSTM is a sequence of hidden states $\mathbf{H}_{1:8} = \{\mathbf{H}_1, \mathbf{H}_2, \dots, \mathbf{H}_8\}$, each carrying information from a specific time step. However, the subsequent Transformer encoder, by itself, is permutation-invariant: it would treat the sequence in any order identically [15]. To preserve temporal order information, we add sinusoidal positional encodings to the sequence.

The Transformer model [15] is a pivotal architecture for sequence modeling. It is composed of stacked encoder layers combining multi-head self-attention, feed-forward blocks, layer normalization, and residual connections. For each position t (where $t = 1, 2, \dots, 8$) and each dimension i (where $i = 0, 1, \dots, 127$), the positional encoding is defined as:

$$\text{PE}_{(t,2i)} = \sin\left(\frac{t}{10000^{2i/d_{\text{model}}}}\right) \quad (9)$$

$$\text{PE}_{(t,2i+1)} = \cos\left(\frac{t}{10000^{2i/d_{\text{model}}}}\right) \quad (10)$$

where $d_{\text{model}} = 256$ is the hidden dimension. This sinusoidal function has the property that PE_{t+k} can be represented as a linear function of PE_t , which helps the model learn relative position information [15]. The positional encoding vectors are added element-wise to the LSTM outputs:

$$\mathbf{H}_t^{\text{enc}} = \mathbf{H}_t + \mathbf{PE}_t \in \mathbb{R}^{256} \quad (11)$$

F. Transformer Encoder with Multi-Head Self-Attention

The positionally encoded sequence $\mathbf{H}_{1:8}^{\text{enc}}$ is then passed through a Transformer encoder consisting of two identical layers [15]. Each layer comprises two subcomponents: a multi-head self-attention mechanism and a position-wise feed-forward network. Residual connections and layer normalization are applied around each subcomponent.

1) *Multi-Head Self-Attention*: The self-attention mechanism allows each position in the sequence to attend to all other positions, computing a weighted combination of information from the entire temporal window. For a single attention head, given query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$ matrices derived from the input sequence, the scaled dot-product attention is computed as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right) \mathbf{V} \quad (12)$$

where $d_k = 64$ is the dimension of each attention head. The scaling factor $\sqrt{d_k}$ prevents the dot products from growing too large in magnitude, which would push the softmax function into regions of extremely small gradients [15].

Instead of performing a single attention function, multi-head attention projects the queries, keys, and values $h = 4$ times with different, learnable linear projections. The attention outputs from each head are concatenated and linearly transformed:

$$\text{head}_i = \text{Attention}(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V) \quad (13)$$

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \dots, \text{head}_4) \mathbf{W}^O \quad (14)$$

where $\mathbf{W}_i^Q, \mathbf{W}_i^K \in \mathbb{R}^{256 \times 64}$, $\mathbf{W}_i^V \in \mathbb{R}^{256 \times 64}$, and $\mathbf{W}^O \in \mathbb{R}^{256 \times 256}$ are learnable weight matrices.

The four attention heads operate in parallel, each learning to focus on different temporal patterns. Based on our analysis, we hypothesize that different heads specialize in distinct aspects of motorcycle dynamics: one head may focus on frames around steering transitions, another on steady-state turning phases, and others on tire slip-related dynamics. This multi-head design enables the network to capture multiple distinct temporal relationships simultaneously.

2) *Position-Wise Feed-Forward Network*: Following the multi-head self-attention mechanism, a position-wise feed-forward network is applied identically to each position:

$$\text{FFN}(\mathbf{T}^{(1)}) = \text{ReLU}(\mathbf{T}^{(1)} \mathbf{W}_1 + \mathbf{b}_1) \mathbf{W}_2 + \mathbf{b}_2 \quad (15)$$

where $\mathbf{T}^{(1)}$ is the output of the attention subcomponent after residual connection and layer normalization, $\mathbf{W}_1 \in \mathbb{R}^{256 \times 512}$, $\mathbf{W}_2 \in \mathbb{R}^{512 \times 256}$, and $\mathbf{b}_1, \mathbf{b}_2$ are bias vectors. The hidden dimension of 512 is four times the model dimension, following the standard Transformer configuration [15].

Each Transformer layer applies the operations:

$$\mathbf{T}^{(1)} = \text{LayerNorm}(\mathbf{H}^{\text{enc}} + \text{MultiHead}(\mathbf{H}^{\text{enc}}, \mathbf{H}^{\text{enc}}, \mathbf{H}^{\text{enc}})) \quad (16)$$

$$\mathbf{T}^{(2)} = \text{LayerNorm}(\mathbf{T}^{(1)} + \text{FFN}(\mathbf{T}^{(1)})) \quad (17)$$

with $\mathbf{T}^{(2)}$ serving as the input to the next layer. After two Transformer layers, we obtain the final encoded sequence $\mathbf{T}_t \in \mathbb{R}^{256}$ for $t = 1, \dots, 8$.

G. Regression Head and Output States

After the Transformer encoder, we apply global average pooling across the time dimension to aggregate information from the entire sequence:

$$\mathbf{t}_{\text{global}} = \frac{1}{8} \sum_{t=1}^8 \mathbf{T}_t \in \mathbb{R}^{256} \quad (18)$$

This operation collapses the sequence into a single vector that summarizes the relevant temporal information.

The aggregated vector then passes through a regression head consisting of two fully connected layers with ReLU activation and dropout for regularization:

$$\mathbf{z}_1 = \text{Dropout}(\text{ReLU}(\mathbf{W}_1 \mathbf{t}_{\text{global}} + \mathbf{b}_1)), \quad \mathbf{W}_1 \in \mathbb{R}^{128 \times 256} \quad (19)$$

$$\mathbf{z}_2 = \text{Dropout}(\text{ReLU}(\mathbf{W}_2 \mathbf{z}_1 + \mathbf{b}_2)), \quad \mathbf{W}_2 \in \mathbb{R}^{64 \times 128} \quad (20)$$

The final output layer maps to the five estimated states:

$$\hat{\mathbf{y}} = \mathbf{W}_{out}\mathbf{z}_2 + \mathbf{b}_{out}, \quad \mathbf{W}_{out} \in \mathbb{R}^{5 \times 64} \quad (21)$$

H. Training Strategy

The model is trained end-to-end by minimizing the mean squared error between the predicted and ground-truth states. The loss function is defined as:

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^5 (y_{i,k} - \hat{y}_{i,k})^2 \quad (22)$$

where N is the batch size, i indexes the sample, and k indexes the five output dimensions. We use the Adam optimizer with an initial learning rate of 10^{-4} . The batch size is set to 32, and training proceeds for a maximum of 30 epochs with early stopping based on validation loss (patience of 5 epochs). Dropout with probability $p = 0.2$ is applied within the fully connected layers and within the Transformer encoder to prevent overfitting.

All training data is generated using the BikeSim high-fidelity motorcycle simulator, as described in Section IV. Synthetic data offers the crucial advantage of providing accurate ground truth labels for all five dynamic states, which would be difficult or impossible to obtain from real-world instrumentation at the required scale.

The complete set of learnable parameters is summarized in Table III-H.

Component	Number of Parameters
ResNet50 backbone	23.5 million
BiLSTM (128+128 units)	860,160
Transformer: Multi-head attention (2 layers)	788,480
Transformer: Feed-forward networks (2 layers)	1,050,624
Positional encoding	0
FC layers (256→128→64→5)	41,477
Total	26.2 million

TABLE I

LEARNABLE PARAMETERS OF THE PROPOSED BiLSTM-TRANSFORMER ARCHITECTURE.

IV. EXPERIMENTAL VALIDATION

Acquiring real-world labeled data for motorcycle dynamics estimation presents significant challenges. Accurate ground truth measurements require expensive instrumentation (differential GNSS-RTK, tactical-grade IMUs, optical sensors) exceeding 50,000 euros, while safety considerations limit data collection during aggressive maneuvers. Consequently, research in this domain has extensively relied on high-fidelity simulation [3], [5].

We employ BikeSim, a multi-body dynamics simulator for two-wheeled vehicles incorporating Pacejka Magic Formula tire models [12]. The simulated motorcycle is the Big Sports model (caster angle 24° , trail 0.0882m, mass 281.3kg). The test course consists of six consecutive 45° arcs (radii decreasing from 300m to 100m) connected by 100 m straight segments, with 3.5m lane width on flat terrain.

Data is collected at 60Hz over 28 runs spanning speeds from 20km/h to 130km/h. Gaussian noise is added to sensor

Speed	Metric	State	Baseline	Proposed	Improvement
30 km/h	RMSE	δ	0.0111°	0.0068°	39% ↓
		v_y	0.0053 m/s	0.0035 m/s	34% ↓
	RMSE%	δ	4.71%	2.89%	39% ↓
		v_y	10.02%	6.61%	34% ↓
	RMSE	Δ_y	0.0067 m	0.0044 m	34% ↓
		α_f	0.0015°	0.0009°	40% ↓
	RMSE%	Δ_y	0.38%	0.25%	34% ↓
		α_f	12.56%	7.54%	40% ↓
	RMSE	α_r	0.0015°	0.0009°	40% ↓
		α_r	3.17%	1.90%	40% ↓
57 km/h	RMSE	δ	0.0017°	0.0009°	47% ↓
		v_y	0.0013 m/s	0.0008 m/s	38% ↓
	RMSE%	δ	1.12%	0.59%	47% ↓
		v_y	3.15%	1.95%	38% ↓
	RMSE	Δ_y	0.0027 m	0.0016 m	41% ↓
		α_f	0.0008°	0.0004°	50% ↓
	RMSE%	Δ_y	0.15%	0.09%	40% ↓
		α_f	1.70%	0.85%	50% ↓
	RMSE	α_r	0.0013°	0.0008°	38% ↓
		α_r	0.79%	0.49%	38% ↓
87 km/h	RMSE	δ	0.0017°	0.0009°	47% ↓
		v_y	0.0021 m/s	0.0013 m/s	38% ↓
	RMSE%	δ	4.52%	2.39%	47% ↓
		v_y	0.69%	0.43%	38% ↓
	RMSE	Δ_y	0.0037 m	0.0022 m	41% ↓
		α_f	0.0012°	0.0007°	42% ↓
	RMSE%	Δ_y	0.21%	0.13%	38% ↓
		α_f	1.87%	1.09%	42% ↓
	RMSE	α_r	0.0014°	0.0009°	36% ↓
		α_r	0.41%	0.26%	37% ↓
100 km/h	RMSE	δ	0.0031°	0.0019°	39% ↓
		v_y	0.0047 m/s	0.0031 m/s	34% ↓
	RMSE%	δ	2.23%	1.36%	39% ↓
		v_y	1.00%	0.66%	34% ↓
	RMSE	Δ_y	0.0083 m	0.0055 m	34% ↓
		α_f	0.0023°	0.0014°	39% ↓
	RMSE%	Δ_y	0.47%	0.31%	34% ↓
		α_f	4.25%	2.59%	39% ↓
	RMSE	α_r	0.0023°	0.0015°	35% ↓
		α_r	0.56%	0.36%	36% ↓

TABLE II

QUANTITATIVE COMPARISON BETWEEN BASELINE AND PROPOSED BiLSTM-TRANSFORMER.

measurements to simulate real-world conditions. The dataset comprises over 200,000 image-state pairs, partitioned into training (70%), validation (15%), and test (15%). Sequences of eight consecutive frames are formed with a sliding window stride of four frames.

A. Quantitative Results

Performance is evaluated using Root Mean Squared Error (RMSE) and RMSE percentage. Table II presents the results for both the baseline unidirectional LSTM architecture (4-frame window) and our proposed BiLSTM-Transformer (8-frame window) across four speed scenarios. Our method consistently outperforms the baseline across all speeds and all five estimated states.

The most significant improvements are observed at 57 km/h and 87 km/h, where steering angle RMSE is reduced by 47%. Slip angle estimation improves by up to 50% at mid-range speeds. The steady-state error previously observed at 30 km/h in the baseline is substantially reduced. Furthermore, the proposed architecture achieves faster inference (6.5 ms per frame vs 8.2 ms), representing a 21% speedup despite a modest 5% increase in parameter count.

B. Discussion and Comparison with Prior Work

The improvements demonstrated by the BiLSTM-Transformer architecture are attributable to three key design differences. First, the extended temporal window (8 frames instead of 4) provides longer context, which is particularly beneficial for low-speed steady turns where the baseline exhibited noticeable steady-state offset. Second, bidirectional processing enables the network to anticipate upcoming maneuvers by incorporating future frames into current state estimation. Third, multi-head self-attention allows selective focusing on the most informative time steps for each output state, effectively learning task-specific temporal weighting.

Compared to the baseline approach [13], our method achieves superior performance on all estimation metrics while maintaining the key advantages of the data-driven paradigm: no prior vehicle model required, low-cost sensors, and real-time performance. The attention mechanism additionally provides interpretability not present in the baseline, with different heads specializing in distinct temporal patterns (steering transitions, steady-state turning, yaw rate peaks, lateral acceleration events).

Several factors continue to constrain real-world deployment. Accurate ground truth acquisition for slip angles and lateral velocity requires instrumentation costing over 50,000 euros, and safety considerations limit data collection during near-limit maneuvers. Moreover, models trained on one motorcycle configuration may not generalize to different mass distributions or tire characteristics. These challenges motivate continued research into domain adaptation and semi-supervised learning techniques that can leverage abundant unlabeled real-world data alongside limited labeled examples.

V. CONCLUSION

In summary, our research endeavors in the field of motorcycle lateral dynamics estimation have led to the development of a BiLSTM-Transformer hybrid architecture that fuses monocular visual data with low-cost inertial measurements. The proposed method estimates five key dynamic states steering angle, lateral velocity, lateral offset, and front and rear tire slip angles without requiring prior knowledge of vehicle physical parameters. The architecture combines a ResNet50 backbone for visual feature extraction, a bidirectional LSTM capturing past and future context over an eight-frame window, and a Transformer encoder with multi-head self-attention.

Experimental validation using the BikeSim simulator across speeds from 20 km/h to 130 km/h demonstrates that the proposed architecture significantly outperforms the baseline unidirectional LSTM. Steering angle RMSE is reduced by up to 47% at mid-range speeds, front slip angle estimation improves by up to 50%, and inference time is 21% faster (6.5 ms per frame). The steady-state error previously observed at low speeds is substantially reduced, and the attention mechanism provides interpretability absent in the baseline.

Limitations include validation conducted exclusively in simulation and the need for generalization to different mo-

torcycle configurations. Future work comprises real-world validation on an instrumented motorcycle, domain adaptation from synthetic to real data, extension to additional states such as roll dynamics, and integration into a closed-loop advanced rider assistance system.

ACKNOWLEDGMENT

This work was supported by the French National Research Agency (ANR) through the eMC2 project.

REFERENCES

- [1] Alrazouk, O., Chellali, A., Hadj-Abdelkader, H., Arioui, H.: Vision-based estimation of motorcycle attitude. *IEEE Robotics and Automation Letters* **8**(9), 5759–5766 (2023)
- [2] Cossalter, V.: *Motorcycle Dynamics*. Lulu, 2nd edn. (2006)
- [3] Damon, P.M., Dabladj, H., Ichalal, D., Nehaoua, L., Arioui, H., Mammar, S.: Lateral motorcycle dynamics and rider action estimation: An ltv unknown input observer approach. In: 2016 IEEE Conference on Control Applications (CCA). pp. 711–716 (2016)
- [4] Damon, P.M., Hadj-Abdelkader, H., Arioui, H., Youcef-Toumi, K.: Image-based lateral position, steering behavior estimation, and road curvature prediction for motorcycles. *IEEE Robotics and Automation Letters* **3**(3), 2694–2701 (2018)
- [5] Fouka, M., Nehaoua, L., Arioui, H.: Motorcycle state estimation and tire cornering stiffness identification applied to road safety: Using observer-based identifiers. *IEEE Transactions on Intelligent Transportation Systems* **23**(7), 7017–7027 (2021)
- [6] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* pp. 770–778 (2016)
- [7] Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Computation* **9**(8), 1735–1780 (1997)
- [8] Ichalal, D., Dabladji, M.E.H., Arioui, H., Mammar, S., Nehaoua, L.: Observer design for motorcycle's lean and steering dynamics estimation: a takagi-sugeno approach. In: 2013 IEEE Conference on Decision and Control (2013)
- [9] K. L. Stolle, A.W., Schmidt, S.: Motorcycle trajectory prediction using long short-term memory networks. In: *Proc. 15th International Motorcycle Conference* (2024)
- [10] Miao, J., et al.: Residual learning towards high-fidelity vehicle dynamics modeling with transformer. *arXiv preprint arXiv:2502.11800* (2025)
- [11] Naweed, A., Blackman, R.: The art of riding safely: A critical examination of advanced rider assistance systems in motorcycle safety discourse. *Transportation Research Part F: Traffic Psychology and Behaviour* **107**, 1–15 (2024)
- [12] Pacejka, H.: *Tire and vehicle dynamics*. Elsevier (2005)
- [13] Qian, Y., Pryde, M., Hadj-Abdelkader, H., Nehaoua, L., Tabia, H., Arioui, H.: Learning-based steering estimation for motorcycles via visual-inertial fusion. In: 2025 Fourteenth International Conference on Image Processing, Theory, Tools & Applications (IPTA). pp. 1–6 (2025), doi: 10.1109/IPTA59101.2025.11222056
- [14] Valeo Group: Valeo and hero motcorp announce partnership to co-develop advanced rider assistance systems (aras) for two-wheelers. Press release (January 2026)
- [15] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. pp. 6000–6010 (2017)