MOTOR: A Multimodal Dataset for Two-Wheeler Rider Behavior Understanding

Varun A. Paturkar[1], Shankar Gangisetty[1], C.V. Jawahar[1]

***Abstract*— Two-wheelers account for a disproportionately high share of road fatalities in the Global South. Research on two-wheeler rider behavior, however, lags far behind four-wheelers, where multimodal datasets have driven major advances in Advanced Driver Assistance Systems (ADAS). To address this gap, we present the MOtorized TwO-wheeler Rider (MOTOR) dataset, the first large-scale, multi-view, multimodal resource dedicated to two-wheelers in dense, unstructured traffic. MOTOR comprises 2,500 sequences (25+ hours of video data) collected from 16 riders and integrates synchronized front, rear, and helmet videos, rider eye-gaze from wearable trackers, on-road audio, and telemetry (GPS, accelerometer, gyroscope). Rich annotations capture traffic context, rider state, 12 riding maneuvers spanning conventional and unconventional behaviors, and legality labels (*Legal*, *Illegal*, *Unspecified*). We benchmark rider behavior recognition and maneuver legality classification using state-of-the-art video action recognition backbones (CNN and Transformer-based), extended with multimodal fusion, and find that combining RGB, gaze, and telemetry consistently yields the best performance. MOTOR thus provides a unique foundation for advancing safety-critical understanding of two-wheeler riding. It offers the research community a benchmark to develop and evaluate models for behavior analysis, legality-aware prediction, and intelligent transportation systems. Dataset and code is available at https://varuniiith.github.io/MOTOR-Dataset/**

I. Introduction

Four-wheelers dominate road transport in the Global North, with cars being the primary mode of transport in regions such as the USA [1] and Germany [2]. In contrast, motorized two-wheelers (motorbikes and scooters) are the predominant means of transport in the Global South, including countries like India [3] and Indonesia [4]. This asymmetry has strongly influenced research priorities and safety technologies. The availability of large-scale multimodal driving datasets, including modalities such as RGB videos, LiDAR, GPS/IMU, vehicle telemetry and driver gaze has fueled breakthroughs in four-wheeler tasks such as object detection [5], [6], [7], [8], semantic segmentation [9], [10], intention prediction [11], [12], and trajectory forecasting [13], [14], [15]. Despite constituting a larger share of on-road vehicles and fatal accidents in the Global South [16], two-wheeler behavior remains underexplored, limiting the development of safety-critical models.

As shown in Fig. 1, two-wheelers are the dominant mode of commute in the Global South, central to commercial activities such as delivery services and ride-sharing. Their widespread use raises critical safety concerns: unlike cars, two-wheelers are more prone to accidents due to *sudden acceleration and braking*, *significant lean angles during maneuvers*, and *minimal structural protection*. These dynamics, coupled with close interactions in dense traffic, introduce risks rarely captured in four-wheeler datasets. This gap underscores the need for dedicated resources to study two-wheeler riding behaviors, a challenge our work directly addresses.

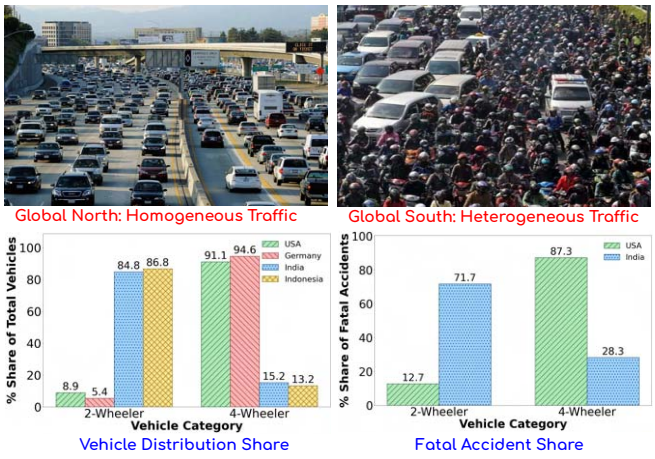

Fig. 1: Comparison of traffic contexts and accident statistics across the Global North and South. *Top Row:* Four-wheelers dominating in the USA vs Two-wheelers in India. *Bottom Row:* Distribution of vehicles (two-wheeler vs four-wheeler) and fatal accidents across North and South.

Recent efforts have begun to address two-wheeler behavior through dedicated datasets. The most notable is RAAD [17], which studied rider intention prediction; myEye2Wheeler [18] introduced rider eye-tracking. As shown in Table I, these datasets are limited in accessibility, usability, details, size, number of views, modalities, conventional, and unconventional riding behaviors.

To address these gaps, we present the MOtorized TwO-wheeler Rider (MOTOR) dataset, the first multi rider, multi-view, multimodal dataset designed specifically for two-wheelers. MOTOR comprises 2,500 sequences collected from 16 riders, integrating synchronized multi-view videos (ego-vehicle front, rear, and helmet views), rider eye-gaze from wearable trackers (Aria [19], Pupil [20]), on-road audio capturing ambient sounds and distractions, and telemetry signals (GPS, accelerometer, gyroscope) for speed, location, and lean angles. The dataset is richly annotated with traffic scene context, rider state, six conventional and six unconventional behaviors, along with their legality labels

[1]CVIT, IIIT-Hyderabad, India. {varuna.paturkar@research., shankar.gangisetty@ihub-data., jawahar@}iiit.ac.in

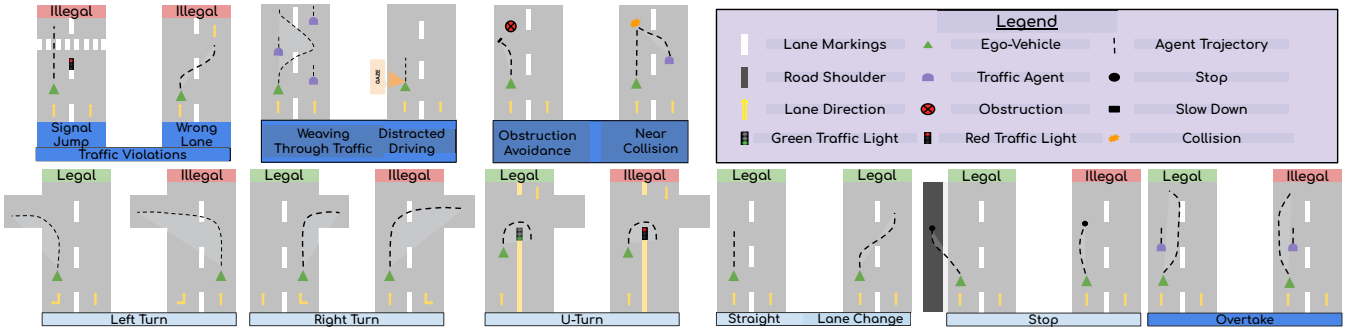

Fig. 2: Illustration of rider behaviors in the MOTOR dataset. Light blue indicates conventional behaviors, dark blue indicates unconventional behaviors. Legal and illegal maneuvers are shown where applicable; behaviors without explicit legality labels are marked as unspecified.

(legal or illegal). As illustrated in Fig. 2, MOTOR captures how riders behave and whether these behaviors comply with traffic rules, offering a unique foundation for safety-critical analysis in two-wheeler research. We benchmark rider behavior recognition on the MOTOR dataset using state-of-the-art action recognition backbones: CNN-based S3D [21], ResNet3D [22], and Transformer-based Video Swin Transformer [23], MViTv2 [24], extended with multi-modal fusion. Beyond behavior classification, we also predict maneuver legality (Legal, Illegal, Unspecified), enabling a more comprehensive understanding of rider actions through visual and non-visual cues.

The main contributions of our work are:

- We introduce **MOTOR**, the first large-scale, multi-rider, multi-view, multimodal dataset for two-wheelers in dense and unstructured traffic. The dataset comprises 2,500 sequences (25+ hours of video data) from 16 riders with synchronized front, rear, and helmet videos, rider eye-gaze, on-road audio, and telemetry. It captures diverse two-wheeler behaviors (see Fig. 2, Fig. 4) including near-collisions, traffic violations, distracted driving, and interactions with vehicles, pedestrians, and street vendors.
- We provide rich annotations of traffic context, rider state, six conventional and six unconventional riding maneuvers, together with legality labels (*Legal*, *Illegal*, *Unspecified*), enabling a legality-aware analysis of how riders behave and comply with traffic rules.
- We benchmark rider behavior recognition and maneuver legality classification using state-of-the-art video action recognition backbones, extended with multimodal fusion of gaze and telemetry, showing that combining modalities consistently improves performance across both tasks compared to video-only baselines.
- We conduct exhaustive experiments to analyze the contribution of each modality (video, gaze, telemetry) and further provide a detailed class-wise accuracy analysis across all backbones (see Fig 8), offering deeper insights into two-wheeler behavior understanding.

II. Related Works

We summarize existing driver (Four-Wheeler) and rider (Two-Wheeler) behavior datasets in Table I.

A. Four-Wheeler Driver Behavior

Driver behavior has been extensively studied in the context of cars through a wide range of datasets [12], [11], [26], [13], [25]. Brain4Cars [12] introduced driver and road-facing videos for maneuver recognition, while HDD [11] extended this with hierarchical annotations of naturalistic driving. The DMD [28] dataset focused more on driver alertness, whereas AIDE [26] captured driver behavior from both inside and outside the vehicle. However, these datasets were primarily collected in structured and sparse traffic. More recent works such as METEOR [13] and IDD-X [25] expanded coverage to heterogeneous and unconstrained road conditions. In parallel, gaze-focused datasets such as DR(eye)VE [29], LBW [30], and DAAD [14] highlighted the importance of attention cues in driving. Together, these efforts showcase the richness and diversity of four-wheeler datasets across structured, unstructured, and gaze-aware contexts.
Nonetheless, they remain centered on cars, with in and out-cabin mounted viewpoints and relatively stable vehicle dynamics. Such setups fail to capture the unique challenges of two-wheelers, including high lean angles during turns, rapid acceleration and braking, and frequent behaviors such as weaving through dense traffic. These limitations underscore the need for dedicated rider-focused datasets.

B. Two-Wheeler Tasks, Datasets, and Methods

Recently, there has been a growing interest in the two-wheeler space, with tasks such as riding pattern recognition [31], riding dynamics analysis [32], safety gear recognition [33], as well as datasets [17], [34], [35], [27], yet the datasets are limited. The RAAD dataset [17] studied rider intention prediction using short clips of six conventional riding maneuvers. While valuable as a first step, RAAD captures only the rider and surrounding vehicles, without incorporating rider gaze. Similarly, myEye2Wheeler [18] introduced an egocentric rider gaze dataset, but its scope is restricted to attention modeling and does not extend to behavioral tasks. Other two-wheeler datasets and efforts remain

TABLE I: Comparison of 4-wheeler and 2-wheeler behavior datasets. Our dataset is unique as it contains multi-modal, multi-view videos from ego-vehicle and helmet, eye gaze, as well as annotated conventional and unconventional behaviors, and legality-related riding scenarios. Note: CRB indicates conventional riding behaviors, and UCRB means unconventional riding behaviors.

Dataset	Ego-vehicle	#Clips	Duration (hrs)	#Views	Resolution	Multi-Modal	Telemetry (gyro,accelero)	Eye Gaze	Audio	CRB	UCRB	Legality (legal,illegal)
IDD-X [25]	Car	3,635	85	2	2560×1440	✗	✗	✗	✗	✓	Partial	✗
HDD [11]	Car	137	104	3	1920×1200	✓	✓	✗	✗	✓	Partial	✗
DAAD [14]	Car	2,028	85	6	1920×1080	✓	✗	✓	✗	✓	✗	✗
Brain4Cars [12]	Car	–	10	2	–	✓	✓	✗	✗	✓	✗	✗
METEOR [13]	Car	1,250	–	2	1920×1080	✓	✗	✗	✗	✓	Partial	✓
AIDE [26]	Car	–	2.4	4	1920×1080	✓	✗	✗	✗	✓	✗	✗
RAAD [17]	2-Wheeler	1,000	–	3	1920×1080	✗	✗	✗	✗	✓	✗	✗
myEye2Wheeler [18]	2-Wheeler	–	100+	1	1920×1080	✓	✓	✓	✗	✗	✗	✗
Oxford RobotCycle [27]	2-Wheeler	–	–	4	–	✓	✓	✗	✗	✓	✗	✗
MOTOR (Ours)	**2-Wheeler**	**2,500**	**25**	**4**	**1920×1080**	✓	✓	✓	✓	✓	✓	✓

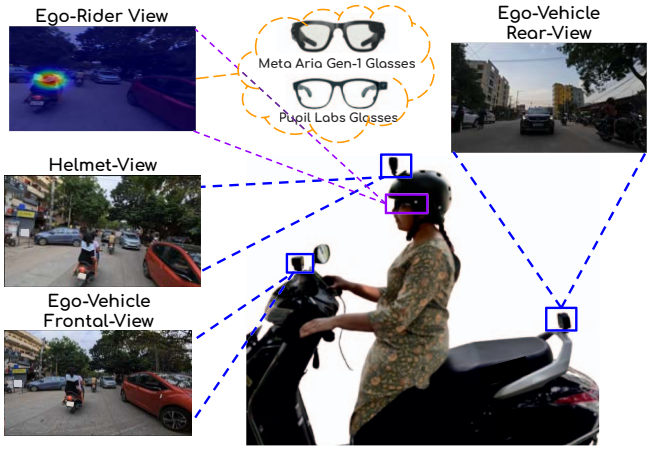

Fig. 3: Data capture setup: Three cameras are oriented towards front-view (ego-vehicle, helmet-mounted), rare-view, and eye-gaze derived from eye-tracking cameras (Aria [19] or Pupil [20] glasses).

task-specific: CDBV [34] collected bike-mounted egocentric views, MoRe [35] targeted motorcycle re-identification, while riding dynamics [32] and powered two-wheeler patterns [31] explored specific motion and traffic patterns. The Oxford RobotCycle [27] examined robotic two-wheeler control, while mobility studies such as EMBARQ [36] highlighted the prevalence of motorized two-wheelers in urban settings. Collectively, these works indicate a growing interest in two-wheeler research but remain narrow in scope. In contrast, our MOTOR dataset introduces a multi-rider, multi-view and multimodal collection with rich annotations covering both conventional and unconventional behaviors along with their legality. This level of diversity opens up a broad research space for advancing rider safety and developing robust models for two-wheeler scenarios.

III. The MOTOR Dataset

In this section, we introduce our dataset and present details of data collection, annotation, and data statistics.

A. Data Collection

Data Capture Platform. The MOTOR dataset was collected using a multi-modal multi-view setup as shown in Fig. 3 with three GoPro-10 cameras and wearable eye-tracking glasses. A front-mounted GoPro, a helmet-mounted GoPro, and a rear-mounted GoPro on the grab rail. Additionally, eye-tracking glasses (Project Aria [19] or Pupil Labs [20]) captured the rider's egocentric view along with gaze information. MOTOR comprises multiple data streams that are synchronized and timestamped.

Video Data. The vehicle was equipped with three GoPro Hero 10 cameras with 1920×1080 resolution, 30 FPS, and video stabilization enabled. The front and rear cameras captured the ego-vehicle's frontal and rear views, providing surrounding traffic context, including side blind spots. The helmet-mounted camera recorded the rider's egocentric view, reflecting the rider's visual perspective of the environment and head movements relevant for rider behavior.

Rider Gaze. Gaze data was recorded using either Aria [19] or Pupil [20] eye-tracking cameras. The Aria device provides eye-tracking at 320×240 resolution and includes an 8 MP RGB camera recording at 1408×1408. In contrast, the Pupil device offers eye-tracking at 192×192 resolution with a 200 Hz sampling rate and a scene camera that captures the rider's egocentric view at 1088×1080.

Audio and Telemetry Data. The Aria [19], Pupil [20] devices, and the GoPro cameras were also used to capture ambient traffic sounds such as horns, engine noise, and rider speech (e.g., with a pillion or during phone calls), which may contribute to riding distractions. Telemetry data was extracted from the GoPro recordings, including GPS signals sampled at 10 Hz for location tracking and inertial measurements (accelerometer and gyroscope) sampled at 200 Hz to characterize rider behavior and ego-vehicle dynamics. This telemetry was synchronized with the video streams using the GoPro telemetry extractor [37].

Data Collection. The dataset comprises 25 hours of riding data collected over 4 weeks, consisting of 25 unique sequences recorded from 16 riders (13 male and 3 female). The diversity of riding data includes varying traffic densities (from peak-hour congestion to sparse early-morning traffic), a wide range of rider experience (2 to 20 years), multiple road types (paved and unpaved, with and without lane markings), and different two-wheeler vehicle types. Importantly, the dataset contains several instances of ego-rider traffic violations, near-collision events, and interactions with pedestrians, street vendors, potholes, and traffic barricades. A comparison of MOTOR with existing two-wheeler and four-wheeler driving datasets is provided in Table I, while Fig. 4 illustrates naturalistic unstructured riding maneuvers leading

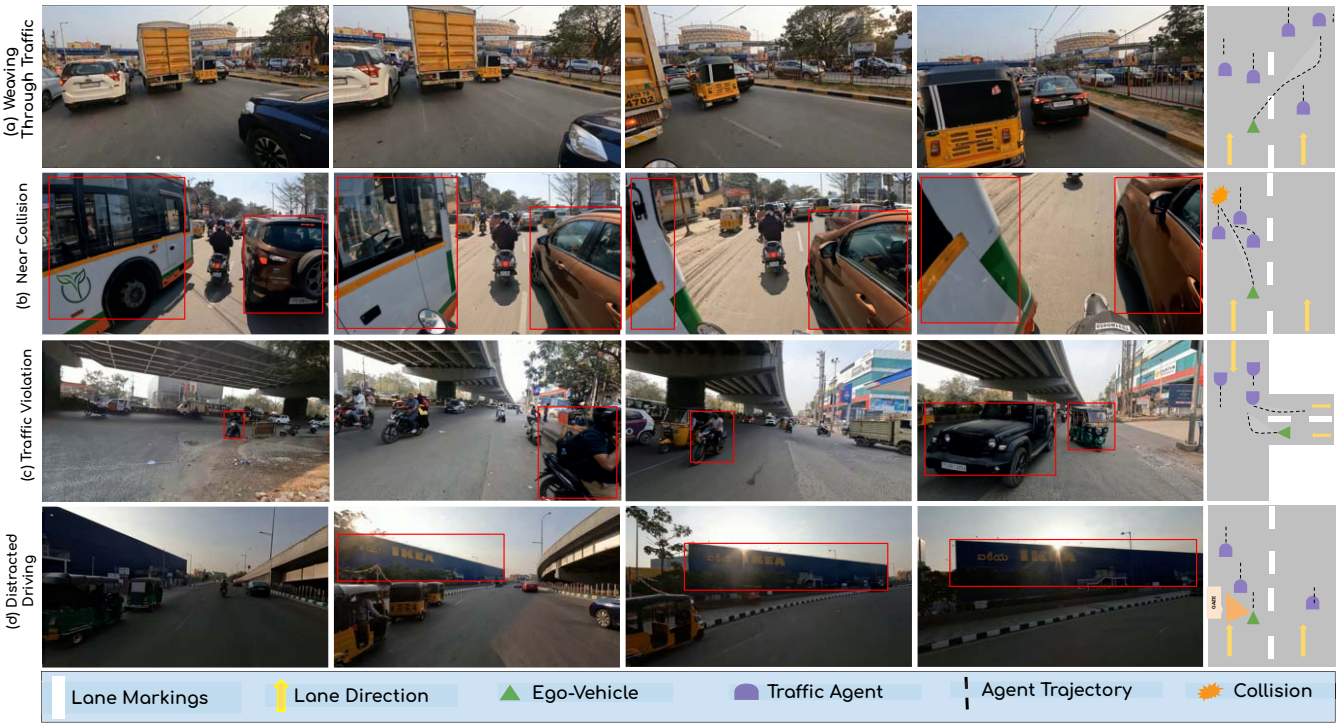

Fig. 4: Data samples helmet-view. *(a)* Ego-rider weaves through dense, slow traffic, overtaking multiple vehicles across lanes. *(b)* Rider squeezes through a narrow gap between a bus and a car, narrowly avoiding the bus. *(c)* Rider rides in the wrong lane against dense oncoming traffic, disrupting flow. *(d)* Rider turns head fully toward a roadside building, diverting gaze from the road amid fast-moving traffic.

to critical situations.

B. Data Annotation and Statistics

To capture a comprehensive view of the traffic scene, ego-rider state, and rider behaviors, the dataset annotations are categorized into four types.

(i) Traffic Scene and Rider State Annotations.
Traffic scene annotations capture the context of ego-rider operation, including time of day, road surface (paved/unpaved), number of lanes, presence of lane markings or dividers, and traffic density. Whereas, *Ego rider state annotations* capture the state of the rider and vehicle, including GPS trajectories, 2D vehicle speeds, rider gaze points, and overall gaze behavior during maneuvers. Gaze behavior is categorized into four classes: *looking straight ahead* (LS), *looking right* (LR), *looking left* (LL), and *glancing sideways* (GS).

(ii) Conventional Rider Behaviors. These annotations capture standard maneuvers performed by the ego-rider, including *Going Straight* (GS), *Left Turn* (LT), *Right Turn* (RT), *Lane Change* (LC), *U-Turn* (UT), and *Stop*.

(iii) Unconventional Rider Behaviors. These annotations capture maneuvers that are potentially dangerous for the rider and should be performed with extreme caution or avoided altogether (see Fig. 4). We categorize them as follows:

- *Overtaking (OT)*: Ego-rider overtakes another traffic agent by accelerating.
- *Weaving Through Traffic (WTT)*: Ego-rider maneuvers through dense traffic by frequently switching between small gaps to move ahead.
- *Obstruction Avoidance (OA)*: Ego-rider swerves or slows down to avoid an obstruction. Sub-categories include avoidance of a vehicle, pedestrian, traffic barricade, or pothole/speed breaker.
- *Distracted Riding (DD)*: Ego-rider diverts attention away from the road ahead, e.g., by looking at a phone, billboard, or other off-road stimuli.
- *Traffic Violations (Vio)*: Instances where the ego-rider commits a violation as defined by the Indian Motor Vehicle Act [38]. Sub-categories include signal jumps, wrong-lane riding, illegal turns and illegal parking.
- *Near Collisions (NC)*: Maneuvers that lead to or narrowly avoid a collision with another traffic agent.

(iv) Legality Annotations. Each maneuver is annotated for its legality based on the Indian Motor Vehicle Act (2017) [38] (see Fig. 5). Legality is categorized into three classes:

- *Legal*: Maneuvers that comply with traffic rules, such as lane following, legal turns, and stopping at signals.
- *Illegal*: Maneuvers that explicitly violate traffic laws, such as signal jumps, wrong-lane riding, illegal turns, or illegal parking.
- *Unspecified*: Maneuvers whose legality depends on context and is not explicitly defined in the Act, such as weaving through traffic or obstruction avoidance.

Data Statistics. Fig. 7 presents an overview of the MOTOR dataset, which includes 1,826 annotated maneuver instances across 12 classes. Maneuver durations range from 2.4 seconds for near-collisions to nearly 14 seconds for

Fig. 5: Data samples of legal and illegal maneuvers. *(a) Legal Left Turn:* The rider correctly takes the left turn by positioning in the left-most lane in advance. *(b) Legal Overtake:* The rider legally overtakes a three-wheeler using the right lane. *(c) Illegal Left Turn:* The rider makes a left turn from the right-most lane, disrupting traffic flow. *(d) Illegal Overtake:* The rider overtakes a three-wheeler from the left within the same lane, even though the right lane is vacant.

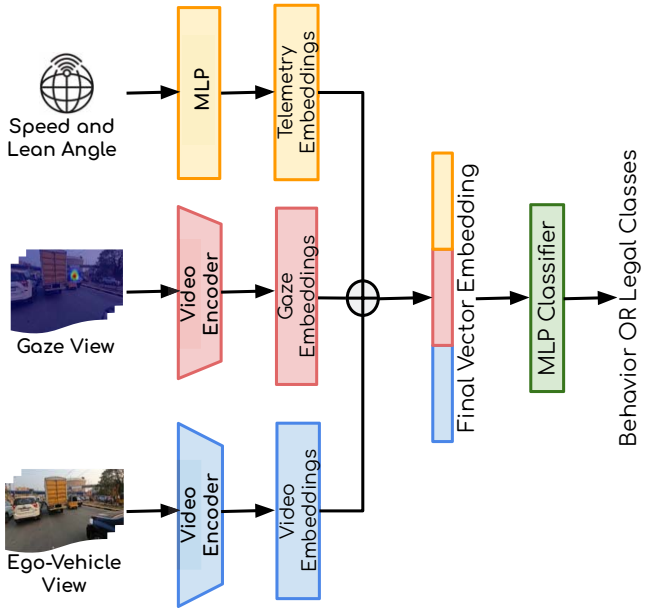

Fig. 6: Our proposed baseline for rider behavior and legality classification. The dual video encoder generates the spatio-temporal features (video and gaze embeddings) along with telemetry features from MLP. These embeddings are concatenated and fed into the MLP classifier to predict the behavioral or legality classes.

traffic violations. Over 60% of instances occur in medium/high traffic, and more than 50% on one- or two-lane city roads (Fig. 7c), reflecting typical urban conditions for two-wheelers. A unique feature is rider head-pose annotations (Fig. 7d), showing how attention shifts with traffic context. Speed and lean angle distributions (Figs. 7e, 7f) reveal strong correlations with maneuvers e.g., high lean during turns and abrupt speed changes in obstruction avoidance or weaving. These statistics highlight the richness of MOTOR in capturing rider state and environmental variability for modeling two-wheeler behavior in dense traffic.

C. Evaluation Tasks and Metrics.

We evaluate two tasks on the MOTOR dataset:

- **Rider Behavior Classification:** Classify rider maneuvers into 11 classes spanning conventional (e.g., turns, lane changes, overtakes) and unconventional (e.g., weaving, obstruction avoidance, violations) behaviors. The *Near Collision* class is excluded due to data sparsity. This task tests a model's ability to capture diverse and fine-grained two-wheeler actions in dense traffic.
- **Maneuver Legality Classification:** Predict whether a rider maneuver is *Legal*, *Illegal*, or *Unspecified*, going beyond behavior recognition to explicitly assess compliance with traffic rules, crucial for safety-critical and traffic-aware systems.

Performance for both tasks was evaluated using *Accuracy* and F_1 scores, reflecting overall correctness and class-balanced performance in skewed label distributions.

Sanity Check. Two professional annotators, trained on maneuver and traffic violation definitions from [38], labeled the dataset under expert supervision. For the first 50 sequences, both annotated independently, after which the expert reviewed their work, resolved discrepancies, and provided feedback. The remaining sequences were also annotated individually, with the expert conducting periodic random checks to ensure accuracy and consistency.

Ethical Statement. All participants receive informed consent prior to data collection, with procedures reviewed and approved by the Institutional Research Board (IRB). Riders were over 18 years of age and held valid driving licenses. Data collection was conducted under naturalistic riding conditions, and no rider was instructed to perform any unconventional or unsafe maneuvers. For safety, instructors supervised the process and participants were briefed on risks before each session. The dataset is anonymized and made publicly available strictly for research purposes.

IV. Baselines

We design a three stream late-fusion architecture, integrating ego-vehicle frontal-view, rider eye-gaze, and vehicle telemetry, as the baseline for rider behavior and legality classification. The overall pipeline is shown in Fig. 6.

Ego-Vehicle Video Stream. The frontal-view video clips are uniformly sampled into 16-frame segments and processed through a video encoder to extract spatio-temporal embeddings. We benchmark four action recognition backbones: CNN-based S3D [21] and ResNet3D [22], and transformer-based Video Swin Transformer [23] and MViTv2 [24].

Gaze Stream. Rider gaze points, recorded using wearable eye-trackers (Aria [19] and Pupil [20]), are used to crop synchronized gaze-centered regions from the video, capturing the rider's localized attention focus. Similar to the video stream, gaze crops are sampled into 16-frame clips and passed through a video encoder, enabling the model

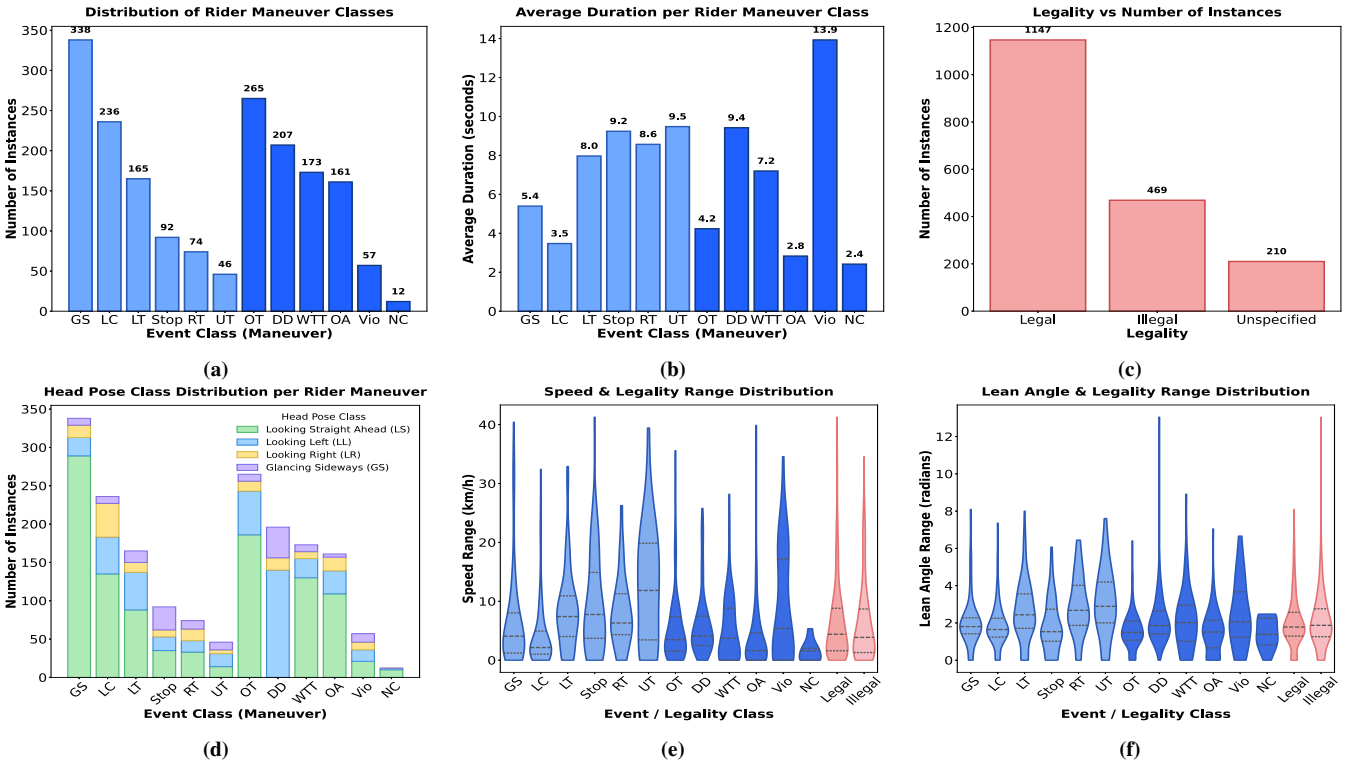

Fig. 7: Annotation Instances and data statistics. *Top row:* Distribution, average duration and legality of rider maneuvers. *Bottom row:* Distribution of head pose class, speed range, and lean angle per rider maneuver class.

to leverage both global scene dynamics and fine-grained attention cues.

Telemetry Stream. Vehicle telemetry (speed and lean angle) provides a compact yet informative representation of motion dynamics. Each signal is resampled into fixed-length sequences of 64 bins, concatenated into a 128-dimensional vector, and projected into an embedding space using a lightweight three-layer MLP with ReLU activations. This stream captures dynamics that are often difficult to infer directly from video, such as subtle lean angles during turns or speed variations during acceleration and braking.

Late Fusion and Classification. The three embeddings from video, gaze, and telemetry are concatenated in a late-fusion scheme to form a unified representation. This fused vector is passed through an MLP classifier consisting of LayerNorm, two hidden layers with ReLU activations and dropout, and a final linear layer, which outputs either the rider behavior or legality classes.

Loss Function. Rider behavior and legality classification face strong class imbalance, dominated by common maneuvers such as *going straight* or *stopping*. We use focal loss [39] with class-balanced weights to address this. For a sample with ground-truth class y and predicted probability p_y, the focal loss is defined as:

$$\mathcal{L}_{\text{focal}} = -\alpha_y (1 - p_y)^{\gamma} \log(p_y),$$

where α_y adjusts for class frequency and γ controls the focusing parameter. This formulation emphasizes on the minority maneuvers, such as weaving through traffic.

V. Experiments

A. Experimental Settings

Implementation Details. All experiments were conducted on a server equipped with an Intel Xeon E5-2640 v4 CPU and four NVIDIA RTX 2080 Ti GPUs. Each of the four video backbones, i.e., S3D [21], ResNet-3D [22], Video Swin Transformer [23], MViTv2 [24] were evaluated under identical conditions to ensure fairness of comparison.

Sampling Strategy. For each video clip, 16 RGB frames were uniformly sampled. Rider gaze crops were synchronized at the same rate to maintain temporal alignment. Telemetry signals were discretized into 64-bin sequences each and concatenated into a 128-dimensional vector per clip, ensuring consistent representation.

Training Procedure. All video backbones were initialized with official PyTorch [40] pretrained weights. All but the final block of each backbone were frozen to reduce computation while enabling task-specific adaptation. The unfrozen block, telemetry encoder and fusion head, were trained end-to-end for 50 epochs with a batch size of 8. We used the AdamW optimizer with cosine annealing, setting learning rates to 1×10^{-4} for the telemetry encoder and fusion head, and 3×10^{-5} for the unfrozen video block, with a weight decay of 1×10^{-4}. Different learning rates were chosen to allow faster convergence for newly initialized modules.

B. Rider Behavior Classification Results

We compare the CNN-based (S3D [21], ResNet3D [22]) and transformer-based (MViTv2 [24], Video Swin Trans-

former (SwinT) [23]) baselines on our MOTOR dataset as reported in Table II. We observe that the transformer-based SwinT [23] baseline outperforms accuracy and F_1 score using all data modalities consistently. SwinT gains over MViTv2 by a margin of 11.4% accuracy and 14.0% F_1 score, showing its effectiveness in modeling complex rider behaviors. Within CNNs, ResNet3D emerges as the stronger architecture, outperforming S3D across all modality settings. A possible reason why MViTv2 under performs compared to ResNet3D and SwinT is its sensitivity to dataset scale and the noise due the the camera movements and vibrations, while the SwinT's hierarchical attention makes it more robust and better suited to dense, noisy scenarios. Table II also reports accuracy and F_1 score across different modalities, underscoring the importance of gaze and telemetry for behavior classification. Using SwinT with all modalities, we achieve accuracy of 52.9%. Removing gaze reduces accuracy by 1.6%, while removing telemetry leads to a 2.6% drop in accuracy. Excluding both causes a 5.2% decline in accuracy, highlighting the effectiveness of full fusion (RGB+Gaze+Telemetry). These results demonstrate that combining attention (gaze) and kinematic (telemetry) cues with visual information significantly improves rider behavior prediction.

TABLE II: Rider Behavior Classification. Comparison of CNN and Transformer-based baselines on MOTOR dataset across different modality combinations.

Baseline	Data Modalities	ACC (↑)	F_1 (↑)	Params (M) (↓)
CNN-based Backbones				
S3D [21]	RGB	38.3	35.3	2.4
	RGB+Gaze	37.3	34.2	4.7
	RGB+Telemetry	39.2	35.8	2.5
	RGB+Gaze+Telemetry	39.3	34.2	4.85
ResNet3D [22]	RGB	48.7	45.4	14.0
	RGB+Gaze	48.2	47.2	28.0
	RGB+Telemetry	48.8	47.1	14.1
	RGB+Gaze+Telemetry	49.1	48.1	28.5
Transformer-based Backbones				
MViTv2 [24]	RGB	32.6	32.4	7.5
	RGB+Gaze	39.4	34.5	15.01
	RGB+Telemetry	39.8	36.1	7.6
	RGB+Gaze+Telemetry	41.5	37.5	15.1
Swin T [23]	RGB	47.7	46.3	7.6
	RGB+Gaze	50.3	46.9	15.1
	RGB+Telemetry	51.3	47.2	7.7
	RGB+Gaze+Telemetry	**52.9**	**51.5**	15.2

C. Maneuver Legality Classification Results

We also benchmark CNN-based models and transformer-based baselines for maneuver legality classification on the MOTOR dataset as shown in Table III. Among all baselines, SwinT [23] achieves the best overall performance, reaching 69.0% accuracy and 53.6% F_1 score with all modalities. Compared to MViTv2, SwinT improves by 4.7% in accuracy, demonstrating stronger robustness in noisy, heterogeneous traffic conditions. Interestingly, within CNNs, S3D proves to be highly competitive: despite having fewer trainable parameters, it outperforms ResNet3D across all modality settings (by 2.1% accuracy and 3.6% F_1 score), and even approaches SwinT's transformer-level performance. This may be attributed to S3D's lightweight 3D convolutions, which capture short-term motion cues crucial for legality assessment, such as detecting wrong-lane usage, illegal turns, or overtakes from the wrong side. Table III also highlights the contribution of different modalities. Full fusion on SwinT yields the strongest performance, while removing gaze reduces accuracy by 4.0% and excluding telemetry leads to a 6.3% drop. Eliminating both modalities results in a 10.6% decline compared to full fusion. These results further highlight the role of gaze as a complementary attention cue and telemetry as a dominant kinematic signal (e.g., speed and lean angle) in shaping accurate legality prediction for two-wheelers and to understand the two-wheeler behavior.

TABLE III: Rider Legality Classification: CNN and Transformer-based baselines on MOTOR dataset across different modality combinations.

Baseline	Data Modalities	ACC (↑)	F_1 (↑)	Params (M) (↓)
CNN-based Backbones				
S3D [21]	RGB	62.9	48.2	2.4
	RGB+Gaze	62.4	48.8	4.7
	RGB+Telemetry	64.5	47.8	2.5
	RGB+Gaze+Telemetry	64.9	51.3	4.8
ResNet3D [22]	RGB	59.6	45.1	14.0
	RGB+Gaze	60.3	45.7	28.0
	RGB+Telemetry	61.8	46.9	14.1
	RGB+Gaze+Telemetry	62.9	47.7	28.5
Transformer-based Backbones				
MViTv2 [24]	RGB	58.2	45.8	7.5
	RGB+Gaze	61.9	46.2	15.0
	RGB+Telemetry	62.6	49.4	7.6
	RGB+Gaze+Telemetry	64.3	52.1	15.1
Swin T [23]	RGB	58.4	47.9	7.6
	RGB + Gaze	62.7	48.5	15.1
	RGB + Telemetry	65.0	53.5	7.7
	RGB+Gaze+Telemetry	**69.0**	**53.6**	15.2

D. Analysis of Confusion Matrices for Rider Maneuver Classification

Fig. 8 presents the confusion matrices for rider maneuver classification across backbones with gaze and telemetry fusion. Among them, the SwinT delivers the most reliable predictions, showing clear separation between *lane change* and *overtake*, *stop* and *going straight*, as well as *left* versus *right turns*. This suggests that its hierarchical attention effectively captures temporal dependencies and contextual cues in dense traffic, making it more robust to noise and ambiguous rider motion patterns. Nevertheless, some confusions persist. *Obstruction avoidance* is frequently misclassified as *lane change* or *overtake*, since all involve lateral deviations. Similarly, *traffic violations* are often misclassified because their recognition depends on scene-level cues (e.g., traffic signals, road markings) that are not always visible in dense, unstructured traffic conditions.

VI. Conclusion and Future Work

In this work, we presented **MOTOR**, the first large-scale, multi-rider, multi-view, and multimodal dataset dedicated to understanding two-wheeler behavior in dense, unstructured traffic. MOTOR integrates synchronized video, rider gaze, audio, and telemetry with rich annotations covering conventional and unconventional maneuvers, along with legality labels. Through extensive experiments and analysis we

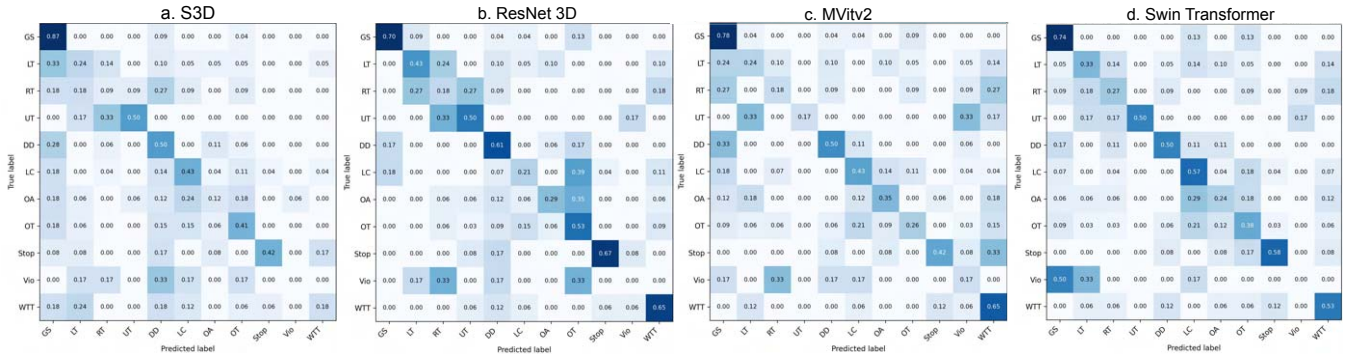

Fig. 8: Confusion matrices for rider behavior classification using S3D [21], ResNet3D [22], MViTv2 [24], and SwinT [23] on the MOTOR dataset.

highlight some key lessons for two-wheeler behavior: gaze provides complementary attention cues, telemetry captures dominant kinematic patterns such as lean and speed, and unconventional behaviors remain challenging to classify due to their variability and overlap with conventional maneuvers. Beyond benchmarking, MOTOR offers a valuable resource for the research community to explore legality-aware modeling and the development of safety-critical applications tailored to two-wheelers. In future, we aim to expand on the dataset with diverse weather and lighting conditions as well as look to improve the baseline architecture by fusing more camera streams and modalities.

VII. Acknowledgment

The project was supported by iHub-Data and Mobility at IIIT Hyderabad. We would also like to thank Shubham Kumar and Md. Azhar for their help with the data collection.

References

[1] U. D. of Transportation Federal Highway Administration, "Highway statistics 2022," 2022. [Online]. Available: https://highways.dot.gov/
[2] Eurostat, "Transport database," 2024. [Online]. Available: https://ec.europa.eu/eurostat/web/transport/database
[3] M. of Road Transport and I. Highways, "Annual report: 2023–2024," 2024. [Online]. Available: https://morth.nic.in/en/annual-report
[4] A. T. Observatory, "E mobility country profile," 2023. [Online]. Available: https://asiantransportobservatory.org/documents/67/Indonesia_20231002b.pdf
[5] H. Caesar *et al.*, "nuscenes: A multimodal dataset for autonomous driving," in *CVPR*, 2020.
[6] P. Sun *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *CVPR*, 2020.
[7] G. Li *et al.*, "Large car-following data based on lyft level-5 open dataset," in *ITSC*, 2023.
[8] F. Yu *et al.*, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *CVPR*, 2020.
[9] G. Varma *et al.*, "Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments," in *WACV*, 2019.
[10] X. Huang *et al.*, "The apolloscape dataset for autonomous driving," in *CVPR Workshops*, 2018.
[11] V. Ramanishka *et al.*, "Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning," in *CVPR*, 2018.
[12] A. Jain *et al.*, "Car that knows before you do: Anticipating maneuvers via learning temporal driving models," in *ICCV*, December 2015.
[13] R. Chandra *et al.*, "Meteor: A dense, heterogeneous, and unstructured traffic dataset with rare behaviors," in *ICRA*, 2023.
[14] A. Wasi *et al.*, "Early anticipation of driving maneuvers," in *ECCV*, 2024.
[15] R. Mahjourian *et al.*, "Unigen: Unified modeling of initial agent states and trajectories for generating autonomous driving scenarios," in *ICRA*, 2024.
[16] MORTH, "Road accidents in india 2022," 2023. [Online]. Available: https://morth.nic.in/road-accident-in-india
[17] S. Gangisetty *et al.*, "Icpr 2024 competition on rider intention prediction," in *ICPR*, 2024.
[18] B. V. Kumar *et al.*, "myeye2wheeler: A two-wheeler indian driver real-world eye-tracking dataset," in *ITSC*, 2024.
[19] J. Engel *et al.*, "Project aria: A new tool for egocentric multi-modal ai research," *arXiv*, 2023.
[20] M. Kassner *et al.*, "Pupil: An open source platform for pervasive eye tracking and mobile gaze-based interaction," *arXiv*, 2014.
[21] S. Xie *et al.*, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *ECCV*, 2018.
[22] D. Tran *et al.*, "A closer look at spatiotemporal convolutions for action recognition," in *CVPR*, 2018.
[23] Z. Liu *et al.*, "Video swin transformer," in *CVPR*, 2022.
[24] Y. Li, C.-Y. Wu, H. Fan, K. Mangalam, B. Xiong, J. Malik, and C. Feichtenhofer, "Mvitv2: Improved multiscale vision transformers for classification and detection," in *CVPR*, 2022.
[25] C. Parikh *et al.*, "Idd-x: A multi-view dataset for ego-relative important object localization and explanation in dense and unstructured traffic," in *ICRA*, 2024.
[26] D. Yang *et al.*, "Aide: A vision-driven multi-view, multi-modal, multi-tasking dataset for assistive driving perception," in *ICCV*, 2023.
[27] E. Panagiotaki *et al.*, "The oxford robotcycle project: A multimodal urban cycling dataset for assessing the safety of vulnerable road users," *IEEE Transactions on Field Robotics*, 2025.
[28] J. D. Ortega *et al.*, "Dmd: A large-scale multi-modal driver monitoring dataset for attention and alertness analysis," in *ECCV*, 2020.
[29] A. Palazzi *et al.*, "Predicting the driver's focus of attention: the dr(eye)ve project," *IEEE TPAMI*, 2019.
[30] I. Kasahara *et al.*, "Look both ways: Self-supervising driver gaze estimation and road scene saliency," in *ECCV*, 2022.
[31] F. Attal *et al.*, "Powered two-wheeler riding pattern recognition using a machine-learning framework," *IEEE Transactions on Intelligent Transportation Systems*, 2015.
[32] M. Bartolozzi *et al.*, "Data-driven methodology for the investigation of riding dynamics: A motorcycle case study," *IEEE Transactions on Intelligent Transportation Systems*, 2023.
[33] J. A. Sanchez-Rodriguez *et al.*, "Motorcycle safety gear recognition with deep learning," in *TEMSCON LATAM*, 2024.
[34] Y. He *et al.*, "Cdbv: A driving dataset with chinese characteristics from a bike view," *IEEE Access*, 2019.
[35] A. Figueiredo *et al.*, "More: A large-scale motorcycle re-identification dataset," in *WACV*, 2021.
[36] EMBARQ, "Motorized two wheelers in indian cities," 2014.
[37] "Telemetry extraction for gopro," 2024. [Online]. Available: https://goprotelemetryextractor.com/telemetry-overlay-gps-video-sensors
[38] MORTH, "Indian motor vehicle driving regulation 2017," 2017. [Online]. Available: https://morth.nic.in/sites/default/files/Motor-Vehicle-Driving-Regulation-2017.pdf
[39] T.-Y. Lin *et al.*, "Focal loss for dense object detection," in *ICCV*, 2017.
[40] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, 2019.