AutoMT: A Multi-Agent LLM Framework for Automated Metamorphic Testing of Autonomous Driving Systems

I. Supplementary material 1

A. *Details of ontology elements*

In this paper, we identified a total of 38 traffic rules from Germany and 72 from California. Due to space limitations, we do not list all the ontology elements in the paper. Instead, we summarize the extracted ontology elements in Tables I and II, corresponding to Germany and California respectively. These tables follow the same subcategory structure as paper. Since the expected behavior of the ego-vehicle remains consistent across all rules, it is omitted from the these two tables for brevity.

Road Type
All roads, Intersection, Port Area, Field path, Forest path, Footpath, Bicycle path, Industrial area, Residential area, Pedestrian zone, Tunnel, Level crossing, Motorway, Hard shoulder
Manipulation (Adds)
St. Andrew's Cross sign, Bus Stop Sign, Start of 30 km/h Zone Sign, Maximum Speed Limit Sign, No Entry Sign, Prohibition for Motor Vehicles Sign, Prohibition for Vehicles of All Kinds Sign, Warning Signs, Right-of-Way Regulation Signs, Stop sign, Yield sign, Pedestrian crossing marking, Lane markings (solid, broken, directional arrows), Traffic island, Reflective posts, Traffic light (Signal system), Guardrail, Chevron signs, Distance markers, Overhead signage, Pedestrian, Bicycle, Bicycle with auxiliary motor, Electric micro-vehicle, Passenger, Driver in front, Vehicle behind, Railway employee with flag, Obstacle on the road, Tractor, Horse-drawn vehicle, Trailer, Animal, Dog, Deer, Child, Blind pedestrian, Person with limited mobility, Skater/scooter rider, Police officer (directing traffic), Construction worker, Overhanging load, Vehicle convoy, Road narrowing, Emergency vehicle, Stopped vehicle, Oncoming vehicle, Turning vehicle, Rail vehicle, Public transport bus, School bus, Line bus, Motor vehicle, Multi-track motor vehicle
Manipulation (Replaces)
Fog, Snowfall, Rain, Black ice, Strong wind, Glare from sun, Poor visibility (general), Night driving (darkness)

TABLE I
Complete ontology elements extracted from German traffic rules

B. *Details of baseline*

Table III illustrates the detailed prompt used in the baseline Auto MT pipeline with LLM-based MR generation without incorporating traffic rules. Table IV illustrates the detailed prompt used in the baseline Auto MT pipeline with LLM-based MR generation with traffic rule guidance.

Road Type
All roads, Intersections, Crosswalks, School Zones, Unmarked Crosswalk, Work Zone, Alleys, On-ramps, Off-ramps, Freeways, Railroad Crossings, Roundabouts
Manipulation (Adds)
Red Light, Yellow Light, Green Light, Red Arrow, Green arrow, STOP Sign, YIELD Sign, 5-Sided Sign, Warning Signs, Crosswalk Markings, Limit lines, Bike Lanes, Center Left Turn Lanes, Flashing Red Light, Flashing Yellow Light, No U-Turn Sign, Wrong Way Sign, Do Not Enter Sign, Speed Limit Signs, High-Occupancy Vehicle (HOV) Lanes, Ramp Meter Signals, Shared Lane Markings (Sharrows), Solid and Broken Yellow Lines (as passing/no-passing indicators), Reflective Road Markers, Railroad Crossing Sign (Crossbuck), Pedestrian Signal Indicators (WALK/DON'T WALK), Raised Pavement Markers, Traffic infrastructure, Vehicle, Emergency vehicle, Tow truck, Road work vehicle, Pedestrian, Person using roller skates, Person using a skateboard, Person with a disability using a wheelchair, Person with a disability using a tricycle, Person with a disability using a quadricycle, Child, Senior (elderly person), Person with small children, Bicyclist, Heavy Traffic, Motorcyclist, Animals, deer, dogs, Traffic Officer, Flaggers (in construction zones), Escorted School Children, School Patrol, Crossing Guard, School bus, collision, Livestock, Obstacle
Manipulation (Replaces)
Rain, Snow, Mud, Ice, Wet road, Fog, Heavy smoke, High winds, Low lighting (implied), Sun Glare, Flooded Roads, Dust Storms, Black Ice, Potholes, Loose Gravel, Drizzle, Slippery road

TABLE II
Complete ontology elements extracted from California traffic rules

Role Setting
You are an expert in traffic rules and scene understanding.
Prompt
Metamorphic Testing (MT) is a method used in autonomous vehicle testing. It defines how the behavior of a system should change when its inputs are changed in a specific way. Given an image of a traffic scene, generate a Metamorphic Relation (MR) in the format: Given the ego-vehicle approaches to [Road Type] When method [Manipulation] Then ego-vehicle should [Ego-Vehicle Expected Behavior] The [Ego-Vehicle Expected Behavior] must be one of the following four actions: - slow down, - keep current, - turn left, - turn right Additional context about the ego-vehicle: {vehicle_info} Example: Given the ego-vehicle approaches to an intersection When method adds a steady red light on the roadside Then ego-vehicle should slow down Now generate an MR for the image below. Only output the MR without explanation User: One Image or Video.

TABLE III
Prompt for Auto MT Pipeline with LLM-based MR Generation **without traffic rule**

We selected nine manually extracted Metamorphic Relations

Role Setting

You are an expert in traffic rules and scene understanding.

Prompt

Metamorphic Testing (MT) is a method used in autonomous vehicle testing.

Traffic rules: It defines how the behavior of a system should change when its inputs are changed in a specific way.

{Traffic Rules}: All traffic rules of this region. e.g. If a pedestrian makes eye contact with you, they are ready to cross the street. Yield to the pedestrian.

Given an image of a traffic scene and traffic rules, extract a Metamorphic Relation (MR) from traffic rule in the format:

Given the ego-vehicle approaches to [Road Type]

When method [Manipulation]

Then ego-vehicle should [Ego-Vehicle Expected Behavior]

The [Ego-Vehicle Expected Behavior] must be one of the following four actions:

- slow down, - keep current, - turn left, - turn right

Additional context about the ego-vehicle: {vehicle_info}

Example:

Given the ego-vehicle approaches to an intersection

When method adds a steady red light on the roadside

Then ego-vehicle should slow down

Now extract an MR from traffic rules for the image below. Only output the MR without explanation.

User: One Image or Video.

TABLE IV
PROMPT FOR AUTO MT PIPELINE WITH LLM-BASED MR GENERATION WITH TRAFFIC RULE

(MRs) from previous research, each with a clear correspondence to formal traffic rules. As these MRs were originally expressed in inconsistent formats, we standardized them into the ”Given–When–Then” structure used in our paper. The table below presents the selected MRs, along with their unified format and references to the original sources.

ID	Metamorphic Relation (Given–When–Then Format)
MR1	Given the ego-vehicle approaches to any roads, When method adds a pedestrian on the roadside, Then ego-vehicle should slow down. [1]
MR2	Given the ego-vehicle approaches to any roads, When method adds a speed limit sign on the roadside, Then ego-vehicle should slow down. [1]
MR3	Given the ego-vehicle approaches to any roads, When method replaces time into night, Then ego-vehicle should slow down. [1]
MR4	Given the ego-vehicle approaches to any roads, When method replaces weather into snow, Then ego-vehicle should slow down. [1]
MR5	Given the ego-vehicle approaches to any roads, When method replaces sunny with rainy, Then ego-vehicle should slow down. [2]
MR6	Given the ego-vehicle approaches to any roads, When method adds a vehicle in front of the ego-vehicle, Then ego-vehicle should slow down. [2]
MR7	Given the ego-vehicle approaches to any roads, When method adds a cyclist in front of the ego-vehicle, Then ego-vehicle should slow down. [3]
MR8	Given the ego-vehicle approaches to any roads, When method adds a red light on the roadside, Then ego-vehicle should slow down. [3]
MR9	Given the ego-vehicle approaches to any roads, When method adds a green light on the roadside, Then ego-vehicle should keep current. [3]

TABLE V
MR OF AUTO MT WITH MANUALLY GENERATED MR

C. Details of ADSs

All autonomous driving systems (ADS) are independently trained on both the German A2D2 dataset and the Californian Udacity dataset for two prediction tasks: steering angle and vehicle speed. All models are implemented in PyTorch and trained using the Adam optimizer with a learning rate of 0.001 for 30 epochs. A batch size of 128 is used during training. The loss function is the Mean Absolute Error (MAE), implemented using the L1Loss criterion. The model checkpoint with the lowest validation loss is selected for evaluation. The detailed configurations of each model are summarized as follows:

- **PilotNet** [4]: The structure of the PilotNet model is represented as {3×160×320, Conv(5×5×3→24, stride=2), Conv(5×5×24→36, stride=2), Conv(5×5×36→48, stride=2), Conv(3×3×48→64), Conv(3×3×64→64), Flatten(→27456), FC(27456→1200), FC(1200→100), FC(100→50), FC(50→10→1)}. The model takes a single RGB image as input and processes it through five convolutional layers to extract visual features, followed by three fully connected layers to reduce the features to 50 dimensions. The final prediction is obtained through a two-layer regressor and outputs a single value for speed or steering.
- **Epoch** [5]: The structure of the Epoch model is represented as {3×160×320, Conv(3×3×3→32)→Conv(3×3×32→64)→Conv(3×3×64 →128)→Conv(3×3×128→256), Flatten(→51200), FC(51200→256), FC(256→1)}. The model takes an RGB image as input and processes it through four convolutional layers with batch normalization, ReLU activation, and 2×2 max pooling. The resulting feature map is flattened and passed through a fully connected layer with 256 units and dropout. Finally, the output is produced through a linear layer that predicts either vehicle speed or steering angle, depending on the task mode.
- **ResNet101-based ADS** [1]: The structure of the ResNet101 model is represented as {3×160×320, ResNet101 (modified input), Flatten→FC(→256), FC(256→1)}. The model takes an RGB image of size 160×320 as input and processes it through a ResNet101 backbone with the original classification head replaced. The final fully connected layer is modified to include batch normalization, dropout (p=0.25), and a linear layer that reduces the extracted feature to 256 dimensions, followed by ReLU activation. A final linear layer maps this to a single output value for speed or steering prediction, depending on the task.
- **VGG16-based ADS** [1]: The structure of the VGG16 model is represented as {3×160×320, VGG16 (features→Conv(512→512))→AvgPool→Flatten(→25088), FC(25088→256), FC(256→1)}. The model takes a 160×320 RGB image as input and processes it through the convolutional backbone of VGG16, followed by an additional 3×3 convolutional layer (512→512) and

average pooling. The resulting feature map is flattened into a 25,088-dimensional vector and passed through a fully connected layer to obtain a 256-dimensional embedding. Finally, a linear layer maps the features to a single output value for speed or steering prediction, depending on the task.

The other two ADSs process multiple consecutive frames:

- **CNN-LSTM** [6]: The model takes a sequence of four RGB images (each of size 3×160×320) as input. Each frame is individually processed through a shared CNN encoder consisting of five convolutional layers: Conv(5×5×3→24, stride=2), Conv(5×5×24→36, stride=2), Conv(5×5×36→48, stride=2), Conv(3×3×48→64, stride=1), and Conv(3×3×64→64, stride=1), each followed by batch normalization and ELU activation. The output feature map is downsampled using adaptive average pooling to a fixed size of 2×2 and flattened to a 256-dimensional vector. These vectors across four time steps form a sequence of shape 4×256, which is fed into an LSTM with 100 hidden units. The last LSTM output is passed through two fully connected layers (100→50→10) and a final regression layer to predict a single value for vehicle speed or steering angle, depending on the task.
- **CNN3D** [6]: The CNN-3D model takes a sequence of four RGB images (each of size 3×160×320) and processes them as a 5D tensor with shape 3×4×160×320. The input is passed through five 3D convolutional layers: Conv3D(3→24, kernel=2×5×5, stride=1×2×2), Conv3D(24→36, 2×5×5, 1×2×2), Conv3D(36→48, 2×5×5, 1×2×2), Conv3D(48→64, 1×3×3), and Conv3D(64→64, 1×3×3), each followed by batch normalization and ELU activation. The resulting feature volume is downsampled using adaptive average pooling to a fixed size of 3×3×3 and flattened to a 1728-dimensional vector. This vector is passed through a fully connected layer (1728→256), followed by a two-layer MLP (256→50→10), and finally mapped to a single output value predicting either vehicle speed or steering angle.

D. *Details of evaluation metrics*

As shown in Table VI and Table VII, the prompts used for **Scenario Alignment** and **Logical Alignment** in the paper are presented respectively.

Role Setting
You are a visual reasoning AI specialized in driving scenes.
Prompt
You will be shown two images. The given scenario description is: {scene_description}. Determine whether **both** images match this scenario. Focus on key driving elements like road types, lane markings, vehicles, traffic signs, and pedestrians. Minor visual differences are acceptable, but both images must clearly represent the described scenario. If the description is generic (e.g., “all roads”), reply “Yes” unless the image is clearly unrelated. Reply only with “Yes” if both images reasonably match the scenario description, or “No” otherwise. Image 1: {original_image} Image 2: {modified_image}

TABLE VI
PROMPT FOR **SCENARIO ALIGNMENT**

Role Setting
You are a reasoning assistant for validating the logical consistency and behavioral correctness of metamorphic relations (MRs) in autonomous driving scenarios.
Prompt 1: Logical Consistency Check
Your task is to determine whether the ego-vehicle’s expected behavior is clearly impossible. Example: - If the MR says “turn left” but there is no turn-related element (e.g., sign, road), respond **“No”**. - If the MR adds a “must turn left” sign and the expected behavior is “turn left”, respond **“Yes”**. Respond only “Yes” or “No” based on the MR. If the answer is “No”, briefly explain why. `MR: {MR}`
Prompt 2: Behavior Rationality Check with Speed Context

You are given a metamorphic relation (MR) and the current ego-vehicle speed.
The expected behavior “slow down” is considered unreasonable if the MR adds a speed limit sign that is higher than or equal to the current speed.
Respond only “Yes” or “No”. If the answer is “No”, briefly explain why.
`MR: {MR}`
`Current speed of the ego-vehicle: {speed} km/h`

TABLE VII
PROMPTS FOR **LOGICAL ALIGNMENT**

E.

REFERENCES

[1] Y. Deng, X. Zheng, T. Zhang, H. Liu, G. Lou, M. Kim, and T. Y. Chen, “A declarative metamorphic testing framework for autonomous driving,” IEEE Transactions on Software Engineering, 2022.

[2] H. Yousefizadeh, S. Gu, L. C. Briand, and A. Nasr, “Using cooperative co-evolutionary search to generate metamorphic test cases for autonomous driving systems,” IEEE Transactions on Software Engineering, 2025.

[3] Z. Yang, S. Huang, T. Bai, Y. Yao, Y. Wang, C. Zheng, and C. Xia, “Metasem: metamorphic testing based on semantic information of autonomous driving scenes,” Software Testing, Verification and Reliability, p. e1878.

[4] M. Bojarski, “End to end learning for self-driving cars,” arXiv preprint arXiv:1604.07316, 2016.

[5] C. Gundling, “cg23,” https://bit.ly/2VZYHGr, 2017.

[6] Z. Lai and T. Bräunl, “End-to-end learning with memory models for complex autonomous driving tasks in indoor environments,” Journal of Intelligent & Robotic Systems, vol. 107, no. 3, p. 37, 2023.