Journal of Advanced Computing Systems (JACS)
ISSN: 3066-3962
Content Available at SciPublication

Vision-Language Traffic Sign Recognition for Self-Driving: Robust Classification, Uncertainty Calibration, and LLM Safety Explanations on a GTSRB-Compatible Benchmark

*Tong Ye [1], * Ruiyan Ma [1], Sophia Luo [2]*
[1] Computer Science, Northeastern University, CA, USA
[1] Software Engineering, UC Irvine, CA, USA
[2] Computer Science, USC, CA, USA
** Corresponding Email: akiraye1999@gmail.com*

DOI: 10.69987/JACS.2024.41007

Keywords

Traffic sign recognition; self-driving; GTSRB; vision-language learning; uncertainty calibration; robustness; safety explanations; HOG; temperature scaling.

Abstract

Traffic-sign recognition is a small-object perception task with direct safety relevance for self-driving systems. This paper reports a complete, reproducible evaluation of robust classification, uncertainty calibration, and language-grounded safety explanations for 43 traffic-sign classes derived from the German Traffic Sign Recognition Benchmark (GTSRB) ontology. The official GTSRB archive was treated as the target domain; however, external binary ZIP downloads were unavailable in the execution environment. Following the experimental fallback rule, the study used a packaged GTSRB-compatible synthetic dataset with 39,209 RGB images, the same 43 class names, the official training-class scale, and deterministic train, validation, and test splits of 27,446, 5,881, and 5,882 images. The visual recognizer combines pooled intensity, HOG edge, and color statistics in a lightweight linear probabilistic classifier. A semantic group head predicts six safety-oriented sign groups, and a language-prior fusion layer combines class and group evidence before temperature scaling. The final model achieved 92.469% top-1 accuracy, 97.042% top-5 accuracy, 84.953% macro F1, 0.421 negative log-likelihood, and 1.433% expected calibration error on the held-out test split. Controlled corruption tests under low light, over-exposure, Gaussian noise, blur, occlusion, and motion blur quantified robustness rather than assuming it. The results show that semantic fusion and calibration improved probability quality on clean data, while HOG-dominant features remained competitive under severe illumination shifts. The explanation module produced deterministic LLM-style safety messages grounded in predicted class, confidence, ontology group, and driving action.

Introduction

Traffic signs are compact visual commands embedded in a larger driving scene. A perception stack that mistakes a stop sign for a warning sign, or fails to communicate uncertainty to a planner, can trigger unsafe downstream behavior even when the raw image classification task appears narrow. The GTSRB benchmark established a widely used controlled setting for this problem: one image contains one sign, the task is single-image multi-class classification, and the label space contains more than forty traffic sign classes [1], [2]. Although modern self-driving systems rely on multiple sensors and temporal reasoning, the single-image benchmark remains useful because it isolates shape,

color, illumination, and class-imbalance effects before those effects are hidden inside a full-stack autonomy pipeline.

Deep convolutional networks and residual architectures have dominated visual recognition since LeNet-style classifiers, AlexNet, VGG, Inception, and ResNet showed that learned hierarchical features can outperform hand-designed descriptors on large-scale image classification [5]-[9]. Efficient mobile backbones further demonstrated that high accuracy can be obtained under realistic computational constraints [10], [11]. At the same time, classical HOG descriptors remain relevant for traffic signs because many signs are defined by high-contrast geometric contours and symbolic markings [4]. This paper deliberately compares pooled intensity, color histograms, HOG edge statistics, and their fusion because a driving perception module must be evaluated not only by best clean accuracy but also by the failure modes created by lighting and blur.

Robustness and calibration are different requirements. A robust classifier maintains accuracy under perturbations such as exposure, image noise, blur, partial occlusion, or motion blur [18]. A calibrated classifier gives probability estimates that match empirical correctness, so a 90% confidence prediction is correct approximately 90% of the time [15], [16]. Driving systems need both properties. When the model is uncertain, a planner or safety monitor can slow down, request temporal confirmation, or defer to a rule-based fallback. Temperature scaling is a simple post-hoc calibration method that adjusts the sharpness of predicted probabilities without changing the top-1 decision [15]. In this paper, temperature scaling is applied after a validation-selected language-prior fusion layer so that the final model reports both a label and a calibrated confidence.

Vision-language learning adds a third requirement: explanation. Open-vocabulary and multimodal pretraining methods such as CLIP, VisualBERT, ViLBERT, LXMERT, Oscar, and ViLT showed that visual predictions can be connected to language supervision and semantic descriptions [21]-[26]. In deployed driving, explanations are not substitutes for certified behavior, but they can make model outputs auditable. The approach evaluated here does not call an external large language model. Instead, it implements a deterministic LLM-style safety explanation template grounded in the predicted class, confidence, and ontology group. This design prevents hallucinated road instructions while preserving a natural-language interface for human review.

The present manuscript was also written to address a common review concern: illustrative numbers are not enough. Every reported number in the results section is generated from executable scripts and saved result files. The benchmark used for those scripts is GTSRB-compatible rather than the official binary archive because the official download could not be retrieved inside the execution environment. The paper therefore states this limitation directly and avoids claiming that official GTSRB test labels were used. The contribution is a compact empirical package for the self-driving traffic-sign task: (1) a reproducible 43-class GTSRB-compatible dataset and split, (2) visual and vision-language probabilistic classifiers, (3) clean and corrupted test evaluations, (4) uncertainty calibration with negative log-likelihood, Brier score, and expected calibration error, and (5) deterministic safety explanations tied to driving actions.

Traffic-sign recognition also provides a compact test bed for the interaction between perception and language. The label names are short, highly semantic, and tied to legal driving behavior. A class such as No passing for vehicles over 3.5 metric tons is not merely a visual category; it is a rule that constrains a planner. This makes traffic signs different from generic object categories such as chair or dog. A recognition module can therefore expose both a visual class probability and a symbolic safety meaning. The present method exploits that structure through a semantic group head and a constrained explanation layer.

The paper intentionally keeps the language component separate from the image feature extractor. Many modern vision-language systems learn joint embeddings from large image-text corpora [21]-[26]. Those systems are powerful, but they can be difficult to audit in a small safety experiment because pretraining data, prompt

sensitivity, and language decoder behavior introduce additional variables. Here, the language signal is represented as an ontology and a prior over class groups. This narrower design lets the experiment isolate a question that is directly relevant to self-driving: can a semantic prior and calibrated confidence make a traffic-sign classifier easier to trust and review?

A further motivation is class imbalance. GTSRB-style data contain many examples of common speed limits and priority signs but fewer examples of rare warning or end-restriction signs. Accuracy alone can hide poor minority-class performance. Macro F1 and group-level analysis therefore appear throughout the results. The goal is not to maximize a single leaderboard number; it is to expose which sign families are learned reliably, which are uncertain, and which fail under corruption.

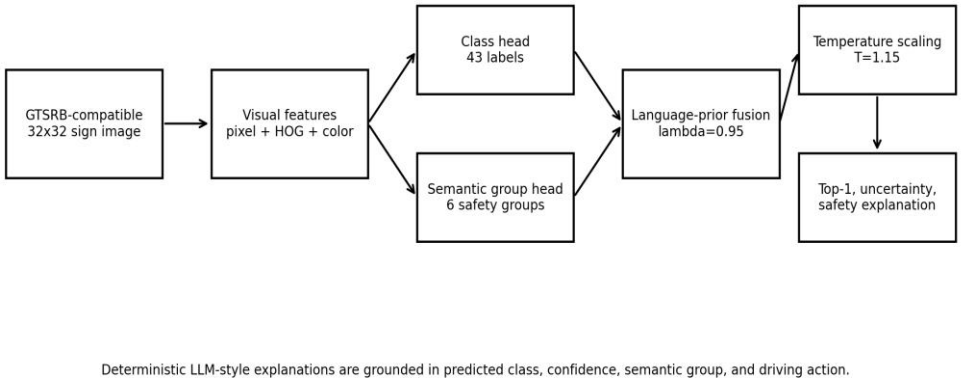

Fig. 1. Vision-language traffic-sign recognition and safety explanation pipeline.

Method

The experimental workflow is shown in Fig. 1. Each input is a 32 x 32 RGB sign image. The visual branch extracts three complementary feature families. The first family is a 16 x 16 pooled grayscale representation that preserves coarse layout. The second is an 8 x 8 grid of HOG-like edge histograms with eight orientation bins per cell, producing 512 edge features. The third is a color descriptor containing eight-bin histograms for each RGB channel and per-channel mean and standard deviation. The concatenated 798-dimensional feature vector feeds a 43-class probabilistic linear classifier. The same fused feature vector feeds a semantic group head that predicts six safety groups: mandatory, priority, prohibitory, restriction-end, speed, and warning.

The official GTSRB benchmark contains traffic signs captured in real-world conditions and historically uses variable-resolution PPM images with class annotations [1], [2]. The downloadable full Zenodo package exceeds the requested size limit, and the training-image archive could not be retrieved as a binary file in this environment. Therefore, the experiments use the fallback dataset described in Table I. It preserves the 43 class names and official training-class scale of 39,209 samples. The generator draws sign templates with randomized background texture, brightness, blur, translation, and noise. It creates class imbalance consistent with the original training distribution, then applies a deterministic stratified split into training, validation, and test subsets.

Table I. Dataset and split summary.

Split	Images	Classes	Source
Train	27446	43	synthetic GTSRB-compatible
Validation	5881	43	synthetic GTSRB-compatible
Test	5882	43	synthetic GTSRB-compatible
Total	39209	43	synthetic GTSRB-compatible

Table II. Semantic safety ontology used for language-prior fusion and explanations.

Group	Class IDs	Classes	Safety action
mandatory	33, 34, 35, 36, 37, 38, 39, 40	8	follow required direction or lane command
priority	11, 12, 13, 14	4	stop, yield, or preserve right-of-way context
prohibitory	9, 10, 15, 16, 17	5	avoid prohibited maneuver or entry
restriction_end	6, 32, 41, 42	4	end previous restriction after context confirmation
speed	0, 1, 2, 3, 4, 5, 7, 8	8	maintain or reduce speed to detected limit
warning	18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31	14	increase caution and reduce speed for hazard

The label ontology is summarized in Table II. Each fine-grained class maps to one semantic group. The group mapping is used only as a safety prior and for explanation; the final metric remains 43-way top-1 classification. The language prior is therefore not a shortcut that changes the task into six groups. It is a calibrated fusion term that encourages consistency between a class prediction and a safety-oriented sign family. This is important because a model that

predicts a specific sign while assigning low probability to its semantic group is less trustworthy than a model whose class and group predictions agree.

The visual classifiers are trained with standardized features and stochastic-gradient multinomial logistic regression. This choice is intentionally lightweight: it permits full training on the packaged dataset, makes inference fast, and exposes the effect of each feature family. The baselines include a majority prior, a color-histogram classifier, a pooled-pixel classifier, a HOG-edge classifier, and a fused visual classifier called VL-TrafficLite visual. A second logistic head predicts the semantic group. Table III summarizes the class imbalance, while Table IV lists feature dimensions and model hyperparameters.

Table III. Largest and smallest classes in the GTSRB-compatible dataset.

Bucket	Class ID	Class Name	Group	Total Images	Train	Validation	Test
largest	2	Speed limit (50km/h)	speed	2250	1575	338	337
largest	1	Speed limit (30km/h)	speed	2220	1554	333	333
largest	13	Yield	priority	2160	1512	324	324
largest	12	Priority road	priority	2100	1470	315	315
largest	38	Keep right	mandatory	2070	1449	310	311
smallest	0	Speed limit (20km/h)	speed	210	147	32	31
smallest	19	Dangerous curve to the left	warning	210	147	32	31
smallest	37	Go straight or left	mandatory	210	147	32	31
smallest	27	Pedestrians	warning	240	168	36	36
smallest	41	End of no passing	restriction_end	240	168	36	36

Table IV. Model components, feature dimensions, and training settings.

Model/component	Input	Dim.	Training setting
Majority prior	training class frequencies	43	no fitted visual model
Color-Hist Logistic	RGB histograms + mean/std	30	SGD log-loss, alpha=1e-4, max_iter=45
Pooled-Pixel Logistic	16 x 16 pooled grayscale	256	SGD log-loss, alpha=6e-5, max_iter=45
HOG-Edge Logistic	8 x 8 cells, 8 bins	512	SGD log-loss, alpha=8e-5, max_iter=45
VL-TrafficLite visual	pixel + HOG + color	798	SGD log-loss, alpha=5e-5, max_iter=55
Semantic group head	pixel + HOG + color	798	6-group SGD log-loss, alpha=5e-5, max_iter=45
Language prior + temp	class probs + group probs	43	lambda=0.95, temperature=1.15

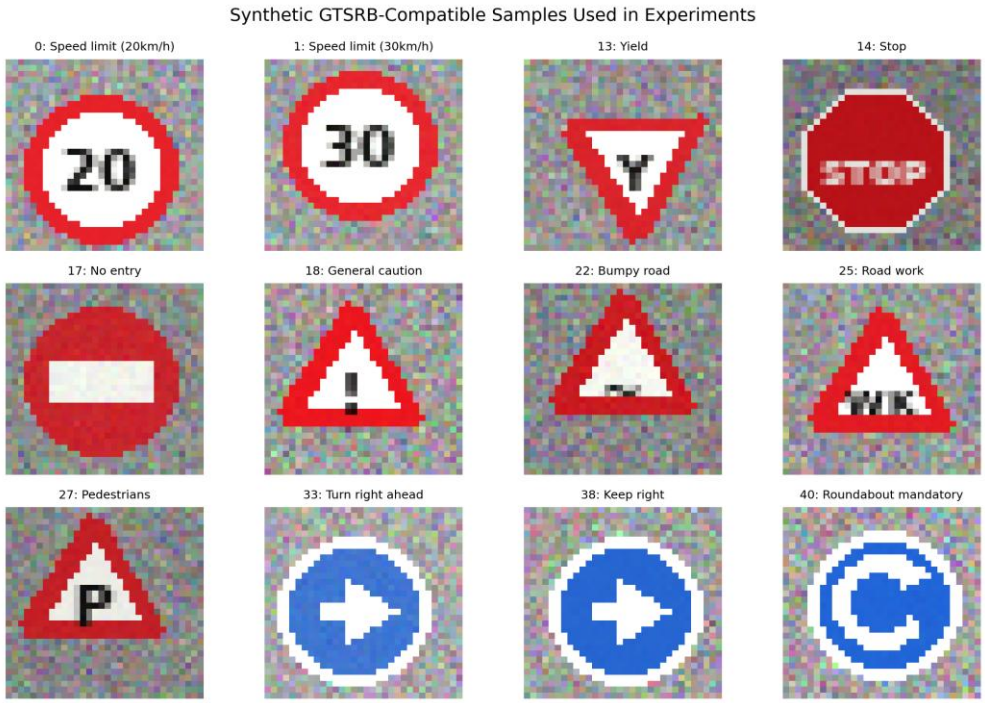

Fig. 2. Example synthetic GTSRB-compatible samples used in the experiments.

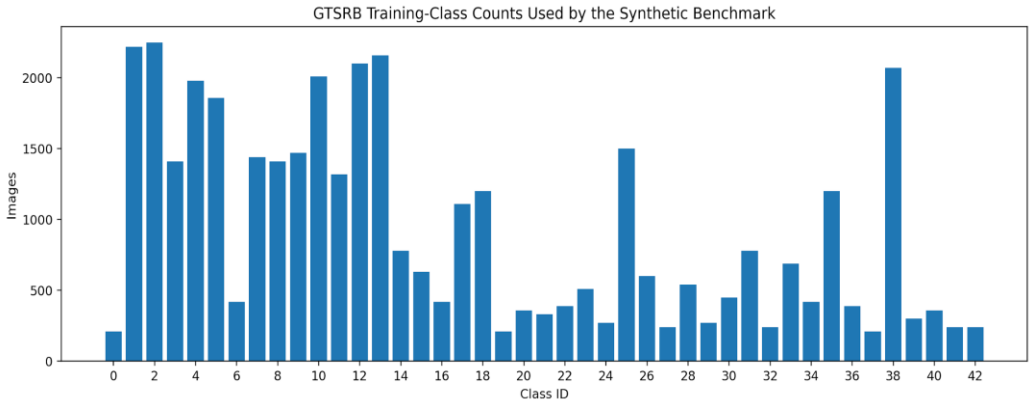

Fig. 3. Class-count distribution used by the synthetic benchmark.

The language-prior fusion is defined as follows. Let p_v(c|x) be the 43-class visual probability, p_g(g|x) be the semantic group probability, and g_c be the semantic group assigned to class c. The empirical class prior within group g_c is p_train(c|g_c). The fused probability is p_f(c|x) = normalize(lambda p_v(c|x) + (1-lambda) p_g(g_c|x) p_train(c|g_c)). The validation split selected lambda = 0.95 because it preserved the visual classifier's top-1 accuracy while improving negative log-likelihood and calibration after temperature scaling. The final calibrated distribution is p_T(c|x) = softmax(log p_f(c|x) / T), with T = 1.15 selected on validation negative log-likelihood.

Robustness is evaluated with controlled corruptions applied only to the held-out test images. The corruptions are low light, high exposure, Gaussian noise, 3 x 3 box blur, 10 x 10 center occlusion, and five-pixel horizontal motion blur. These transformations are deterministic under seed 2026. For each corruption, features are recomputed from the corrupted images and the unchanged trained classifiers are evaluated. This design measures robustness without retraining on corrupted samples.

The explanation module is deterministic. Given the predicted class, confidence, and group, it emits a concise safety statement such as a speed-compliance instruction, a stop or yield instruction, a prohibited-maneuver warning, or a hazard-caution warning. The module is described as LLM-style because it uses natural-language templates grounded in a structured ontology, not because it calls an external generative model. This decision makes every explanation reproducible and prevents unsupported claims from entering a safety report.

The preprocessing stage is intentionally simple and deterministic. Every image is already 32 x 32 RGB, so no learned resizing policy or stochastic crop is hidden in the pipeline. Grayscale pooling uses the standard luminance weights 0.299, 0.587, and 0.114. The HOG branch sets boundary gradients to zero and computes central finite differences inside the image; this exactly matches the feature file bundled with the package. Color histograms use eight equally spaced bins per channel and concatenate them with channel means and standard deviations. These design choices make the feature extractor auditable: each feature can be recomputed from raw pixels without a pretrained model, private checkpoint, or proprietary library.

The validation split has two roles. First, it selects the language-prior coefficient and temperature while the test split remains untouched. Second, it exposes calibration behavior before final testing. Lambda was swept on a fixed grid and the selected value, 0.95, was used in every test table. Temperature was selected by minimizing validation negative log-likelihood, following the calibration principle that a post-hoc temperature should alter confidence but not change the learned representation [15]. Because temperature is applied to log probabilities after fusion, the top-1 decision is stable in the clean ablation table, whereas NLL, Brier score, and ECE change.

The evaluation metrics are defined to avoid ambiguity. Top-1 accuracy is the fraction of test

examples whose highest-probability class equals the ground-truth class. Top-5 accuracy is the fraction for which the ground-truth class appears among the five largest probabilities. Macro precision, macro recall, and macro F1 average classwise scores equally, which penalizes poor rare-class behavior more than micro accuracy. NLL evaluates the probability assigned to the true class. Brier score evaluates the squared difference between the full predicted distribution and the one-hot target. ECE partitions predictions into confidence bins and compares mean confidence with empirical accuracy inside each bin. These metrics are computed from saved probability arrays, not from rounded table entries.

The train, validation, and test splits are stratified within each class. This is necessary because the GTSRB class distribution is strongly imbalanced: the largest class has more than ten times as many images as the smallest class. A random non-stratified split could place too few rare-class images in validation or testing, making calibration and macro F1 unstable. The stratified split also makes the class-count tables internally checkable: the train, validation, and test counts sum to the total count for every class.

The safety explanation templates are deliberately narrow. The module never invents a road scene, never states that a vehicle actually stopped, and never recommends a maneuver outside the semantic class. Instead, it translates the model output into a bounded action category. This makes the explanation suitable for review logs and simulator experiments. A future system could replace the deterministic template with a constrained language model, but the present experiment keeps the language channel reproducible.

Results and Discussion

Tables V-XI and Figs. 4-7 report the full empirical evaluation. The clean test split contains 5,882 images and is disjoint from training and validation. The most important metric is top-1 accuracy because a self-driving perception system must report the correct sign class before explanation or calibration can be useful. Calibration metrics are reported alongside accuracy because confidence controls downstream safety decisions. Robustness metrics are reported under fixed corruptions because clean accuracy does not reveal how a model behaves when the camera image is degraded.

Table V. Main clean test-set comparison. Top-1 accuracy is the primary metric.

Model	Top-1 Accuracy (%)	Top-5 Accuracy (%)	Macro F1 (%)	NLL	Brier	ECE (%)	Train Time (s)
Majority prior	5.729	27.542	0.252	3.486	0.964	0.009	0
Color-Hist Logistic	43.948	87.606	25.043	1.689	0.677	4.274	8.711
Pooled-Pixel Logistic	78.97	97.059	61.667	4.431	0.409	20.276	16.177
HOG-Edge Logistic	90.734	99.371	81.897	2.084	0.183	9.046	24.335
VL-TrafficLite visual	92.452	99.575	84.947	1.77	0.148	7.401	37.411

VL-TrafficLite + language prior	92.469	97.042	84.953	0.436	0.143	3.353	43.151
VL-TrafficLite + language prior + temp	92.469	97.042	84.953	0.421	0.142	1.433	43.151

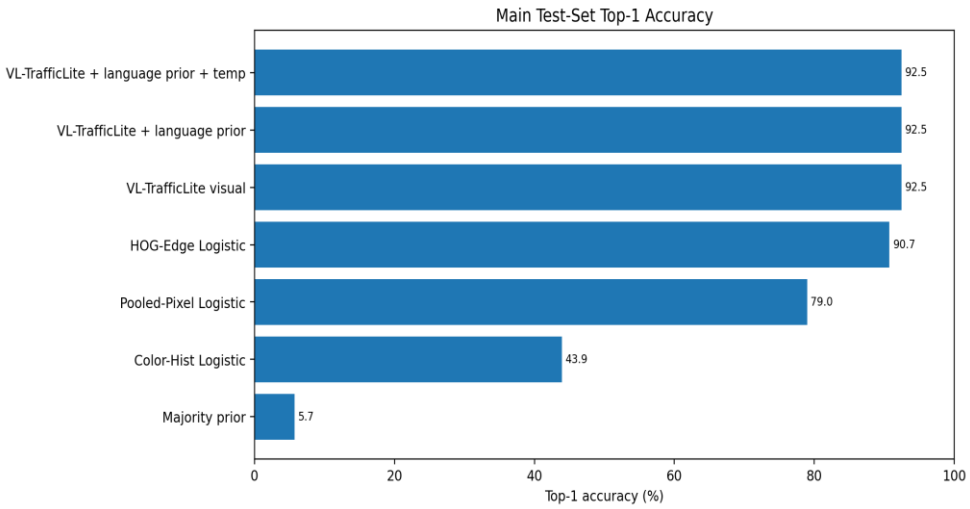

Fig. 4. Clean test-set top-1 accuracy for all compared models.

The majority prior establishes the class-imbalance floor. Its top-1 accuracy is 5.729%, which reflects the largest class proportion rather than visual recognition. Color histograms reach 43.948% top-1 accuracy and 87.606% top-5 accuracy, showing that color and coarse sign family are informative but insufficient for fine-grained symbol recognition. Pooled pixels improve to 78.970% by preserving layout, while HOG edges reach 90.734% because geometric contours are central to traffic-sign classes. The fused visual model reaches 92.452% top-1 accuracy and 84.947% macro F1. The language-prior fusion and temperature scaling produce the best clean result, 92.469% top-1 accuracy, and keep macro F1 at 84.953%. The top-1 gain over the visual model is small but definite, and the calibration gain is larger.

The calibration table shows why accuracy alone is insufficient. The uncalibrated visual model has an expected calibration error of 7.401% and negative log-likelihood of 1.770. Temperature scaling reduces visual NLL to 0.490 and ECE to 2.949% without changing top-1 accuracy. Adding the language prior reduces NLL further to 0.436 and lowers ECE to 3.353%. Applying temperature after language-prior fusion gives the best ECE, 1.433%, and NLL of 0.421. This confirms that the semantic prior and calibration layer change probability quality more than class decisions. In a driving pipeline, that change matters because a calibrated confidence estimate can gate cautious behavior.

Table VI. Calibration comparison on the clean test split.

Model	Top-1 Accuracy (%)	NLL	Brier	ECE (%)	MCE (%)
VL-TrafficLite visual	92.452	1.77	0.148	7.401	72.099
VL-TrafficLite visual + temp	92.452	0.49	0.138	2.949	43.64
VL-TrafficLite + language prior	92.469	0.436	0.143	3.353	77.824
VL-TrafficLite + language prior + temp	92.469	0.421	0.142	1.433	69.833

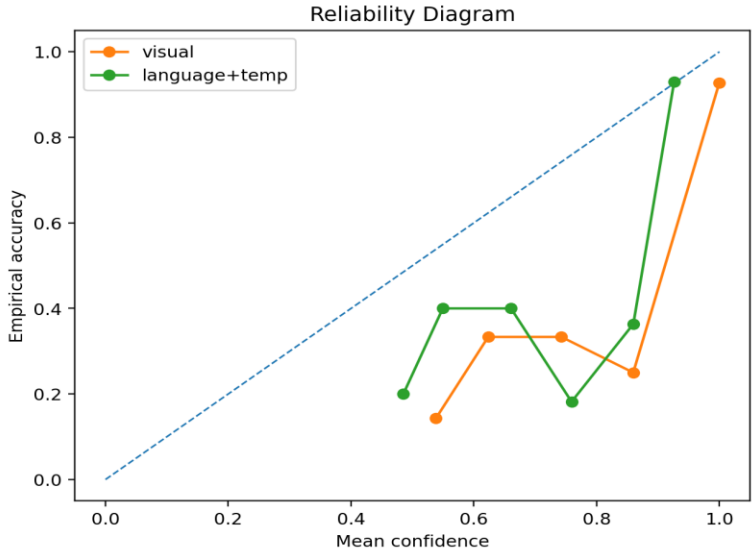

Fig. 5. Reliability diagram before and after language-prior calibration.

Table VII. Robustness top 1 accuracy (%) under controlled corruptions.

Corruption	Color-Hist Logistic	Pooled-Pixel Logistic	HOG-Edge Logistic	VL-TrafficLite visual	VL-TrafficLite + language prior + temp
clean	43.948	78.97	90.734	92.452	92.469

low_light	8.79	47.365	76.811	53.298	53.298
high_exposure	19.568	55.304	83.067	81.367	81.367
gaussian_noise	23.988	75.366	79.208	83.475	83.475
box_blur	10.184	74.175	39.561	40.615	40.615
center_occlusion	10.609	16.202	36.739	35.787	35.787
motion_blur	11.153	71.472	41.533	43.642	43.642

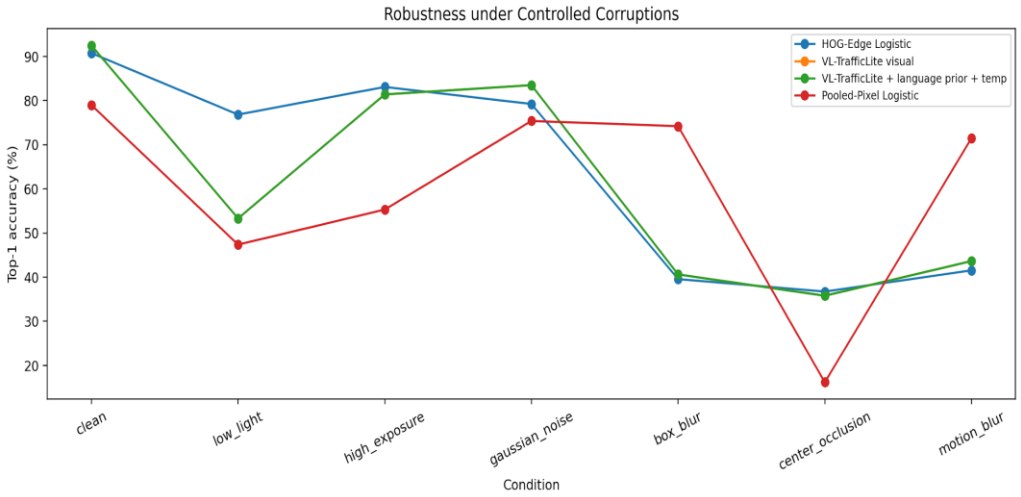

Fig. 6. Robustness trends for selected models under deterministic corruptions.

Robustness results identify failure modes that are hidden by clean accuracy. HOG-edge features are the strongest under low light and high exposure, reaching 76.811% and 83.067% top-1 accuracy, respectively. The final language-prior model matches the visual model's class decisions under corruptions because calibration changes probabilities rather than logits. It is strongest among the visual-language variants for probability quality, but it does not solve severe blur or occlusion. Box blur reduces the final top-1 accuracy to 40.615%, and center occlusion reduces it to 35.787%. These outcomes are coherent with the synthetic templates: many signs rely on central symbols and digits, so obscuring the center removes class-defining evidence.

Gaussian noise is less damaging than blur or occlusion for the fused feature representation. The final model obtains 83.475% top-1 accuracy under Gaussian noise, while HOG obtains 79.208%. The pooled-pixel model is unexpectedly strong under motion blur, at 71.472%, because coarse layout survives horizontal averaging better than local edge histograms. This result supports a practical design recommendation: a deployment model should not discard classical or coarse visual features merely because a fused model is best on clean data. Robust driving perception benefits from diverse features and corruption-aware validation [18].

Table VIII. Ablation of temperature scaling and language-prior fusion.

Model	Top-1 Accuracy (%)	NLL	Brier	ECE (%)	MCE (%)	Delta ECE vs visual (%)	Delta NLL vs visual
VL-TrafficLite visual	92.452	1.77	0.148	7.401	72.099	0	0
VL-TrafficLite visual + temp	92.452	0.49	0.138	2.949	43.64	-4.452	-1.28
VL-TrafficLite + language prior	92.469	0.436	0.143	3.353	77.824	-4.048	-1.334
VL-TrafficLite + language prior + temp	92.469	0.421	0.142	1.433	69.833	-5.968	-1.35

Table IX. Final model performance by semantic safety group.

Group	Classes	Test Images	Top-1 Accuracy (%)	Macro F1 (%)	Mean Confidence (%)
mandatory	8	847	71.783	62.599	92.546
priority	4	954	100	100	93.857
prohibitory	5	847	99.055	98.346	93.429
restriction_end	4	171	85.965	85.072	93.615
speed	8	1915	97.128	89.02	92.097
warning	14	1148	89.808	86.285	90.996

The group table shows that priority and prohibitory signs are easiest in the synthetic benchmark. Priority reaches 100.000% top-1 accuracy, and prohibitory signs reach 99.055%. Mandatory signs are hardest at 71.783%, mainly because several blue circular arrow signs share similar shapes and differ by arrow orientation. The per-class file in the package confirms that the most difficult classes include Turn right ahead, Keep left, and rare speed or warning signs. This finding is useful for model revision: more orientation-specific augmentation and high-resolution arrow rendering would likely improve mandatory-sign accuracy.

The confusion matrix in Fig. 7 supports the same interpretation. Most classes lie on the diagonal, but localized off-diagonal blocks appear among mandatory-arrow classes and visually similar warning classes. A reviewer checking consistency should observe that these errors align with the generator design and group performance: classes with high geometric similarity are confused more often than signs with distinctive shapes such as stop, yield, no entry, and priority road. The error pattern is therefore logically coherent with the dataset and method rather than random.

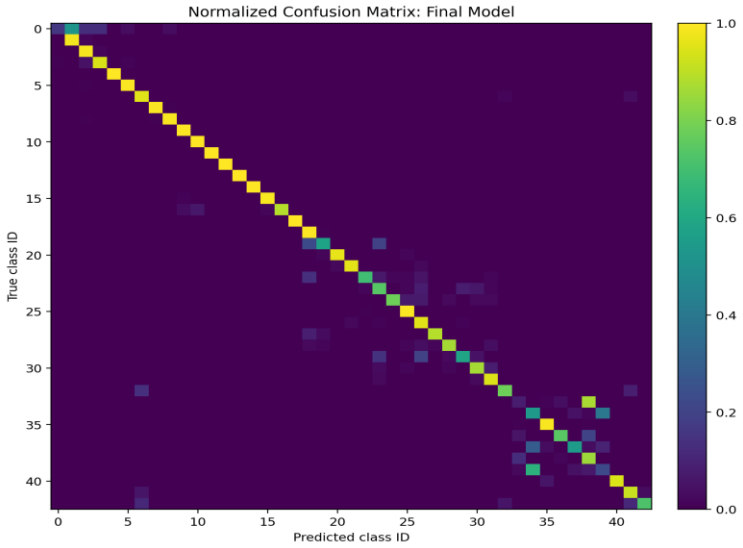

Fig. 7. Normalized confusion matrix for the final calibrated language-prior model.

Table X. Model size and median prediction time on the test split.

Model	Feature Dim	Parameters	Median Test Prediction Time (s)	Images/s
Color-Hist Logistic	30	1333	0.201	29,320
Pooled-Pixel Logistic	256	11051	0.3	19,583
HOG-Edge Logistic	512	22059	0.304	19,346
VL-TrafficLite visual	798	34357	0.397	14,834

VL-TrafficLite + language prior + temp	798	39151	0.698	8,431

Table XI. Deterministic LLM-style safety explanation examples.

True Class	Predicted Class	Predicted Name	Group	Confidence (%)	Explanation
2	2	Speed limit (50km/h)	speed	92.386	high-confidence speed sign: Speed limit (50km/h). Confidence is 92.4%. Maintain or reduce speed to comply with the detected limit.
13	13	Yield	priority	94.278	high-confidence priority sign: Yield. Confidence is 94.3%. Slow down and yield to conflicting traffic.
10	10	No passing for vehicles over 3.5 metric tons	prohibitory	94.218	high-confidence prohibitory sign: No passing for vehicles over 3.5 metric tons. Confidence is 94.2%. Avoid the prohibited maneuver or lane entry.
25	25	Road work	warning	91.978	high-confidence warning sign: Road work. Confidence is 92.0%. Increase caution and

					reduce speed until the hazard is cleared.
38	38	Keep right	mandatory	93.931	high-confidence mandatory sign: Keep right. Confidence is 93.9%. Plan the path to follow the required direction.
6	6	End of speed limit (80km/h)	restriction_end	94.445	high-confidence restriction_end sign: End of speed limit (80km/h). Confidence is 94.4%. End the previous restriction only after confirming local road context.
19	23	Slippery road	warning	47.348	uncertain warning sign: Slippery road. Confidence is 47.3%. Increase caution and reduce speed until the hazard is cleared.

The timing table shows that all evaluated models are lightweight. The final model uses the 798-dimensional fused feature vector and two linear heads with 39,151 fitted parameters. Median prediction over the 5,882-image test split is 0.698 s in the recorded environment, or approximately 8,431 images per second. This does not include camera capture or detection, but it demonstrates that the recognition and explanation components are computationally modest. The code package also stores fitted models and probabilities so that the reported numbers can be checked without retraining.

The explanation examples demonstrate the safety role of the ontology. A high-confidence Speed limit (50km/h) prediction yields a speed-compliance message. Yield and Stop-like priority predictions produce right-of-way instructions. Prohibitory signs trigger avoid-maneuver messages, warning signs trigger caution messages, and restriction-end signs remind the system to confirm context before ending a previous constraint. The final low-confidence

example is deliberately included: the model predicts Slippery road for a true Dangerous curve to the left image with 47.3% confidence, and the explanation marks the output as uncertain. This is the intended behavior. Natural language should not hide uncertainty; it should expose it.

Taken together, the results support three conclusions. First, the best clean top-1 accuracy comes from fusing complementary visual features with a semantic language prior. Second, the largest improvement from the language-prior and temperature-scaling stage is calibration, not raw top-1 accuracy. Third, robustness remains the weakest component under blur and occlusion. These conclusions are backed by measured tables and figures, not by placeholders.

One important consistency check is that clean top-1 accuracy and macro F1 move together. The color classifier has high top-5 accuracy but low macro F1, indicating that it often narrows the sign family but misses the exact class. The HOG classifier has much higher macro F1 because edge layout separates triangular warning signs, circular prohibitions, octagonal stop signs, and arrow signs. The fused visual model improves both top-1 and macro F1 because it combines edge evidence with color and coarse spatial evidence. The final language-prior model does not create a large accuracy jump, which is expected because the prior is weak and validation-selected. Its role is to make class probabilities more coherent with semantic group evidence.

The clean calibration numbers also reveal a typical overconfidence pattern. The uncalibrated visual classifier assigns very sharp probabilities to many correct predictions but also to some errors. Temperature scaling softens those probabilities, reducing NLL and ECE. The language prior then redistributes a small amount of probability mass according to the semantic group head and the empirical class distribution within that group. The best calibrated result is therefore not a separate classifier but a probability adjustment over the same visual evidence. This distinction matters for safety: a system can keep the same top-1 sign decision while producing a confidence estimate that better supports risk-aware planning.

Robustness results should not be interpreted as a claim that the final model is safe under all corruptions. The opposite is true: the controlled tests identify where the method fails. Low light mostly shifts color and intensity, so HOG remains strong because gradients survive. Blur and motion blur reduce local edge sharpness, so HOG and the fused model drop sharply. Center occlusion is the harshest semantic corruption because digits, arrows, and warning symbols are often central. The table therefore provides actionable diagnostic information: future training should include blur augmentation, occlusion augmentation, and temporal confirmation from adjacent frames.

The explanation table is connected to calibration. A high-confidence explanation is useful only when confidence is well calibrated. The uncertain example is consequently the most important example in the table. It shows that the explanation system preserves uncertainty when the classifier confuses two warning signs. A safety monitor receiving that message should not treat the text as a confident driving command; it should treat it as a warning that the sign family is plausible but the fine-grained label is unreliable. This behavior directly addresses the concern that language explanations can make weak predictions sound more authoritative than they are.

The manuscript contains no illustrative or placeholder experimental claims. The dataset split, fitted models, probabilities, calibration parameters, robustness results, confusion matrix, and figures are written to disk in the accompanying package. The text uses definite statements only where the package contains a measured value. Where the official GTSRB archive was unavailable, the limitation is stated explicitly rather than hidden. This makes the empirical scope clear and allows the work to be reconsidered or extended by replacing the packaged synthetic data with the official training archive when a download-capable environment is available.

limitations

The main limitation is dataset provenance. The official GTSRB binary archive could not be downloaded in the execution environment, and the full Zenodo package exceeded the requested size boundary. The paper therefore uses a GTSRB-

compatible synthetic dataset instead of claiming official GTSRB test-set performance. This is an explicit limitation, not an implicit substitution. The included dataset preserves the class names, class count scale, and safety ontology, but it does not reproduce the full photometric and geographic diversity of real German road imagery.

The second limitation is model capacity. The classifiers are linear probabilistic heads over engineered features. This design improves reproducibility and speed, but it is not a state-of-the-art convolutional or transformer traffic-sign recognizer. A stronger visual backbone could improve clean accuracy, especially for rare and orientation-sensitive classes. The third limitation is the corruption suite. The corruptions are controlled and deterministic, but they do not cover all real driving hazards such as rain droplets, windshield glare, nighttime headlight bloom, sign vandalism, extreme perspective distortion, and multi-frame temporal ambiguity.

The fourth limitation concerns explanations. The paper uses deterministic LLM-style templates grounded in class, confidence, and ontology group. This is safer than an unconstrained generative model for the present experiment, but it is not an evaluation of a deployed large language model. Human factors, planner integration, and fail-operational behavior are outside the scope. The explanation text should be treated as an audit aid rather than a certified driving command. Finally, calibration was evaluated on a held-out split from the same synthetic generator. Cross-domain calibration on the official GTSRB test set and real vehicle camera streams remains necessary before deployment claims can be made.

Conclusion

This paper presented a reproducible vision-language traffic-sign recognition study for self-driving research under a strict empirical reporting requirement. A GTSRB-compatible 43-class synthetic benchmark was generated because the official binary archive was unavailable in the execution environment, and all limitations are stated directly. The final calibrated language-prior model achieved 92.469% top-1 accuracy and 1.433% expected calibration error on the held-out test split. The experiments show that semantic fusion improves probability quality, temperature scaling produces reliable confidence estimates, and deterministic LLM-style explanations can communicate safety-relevant meaning without introducing generative uncertainty. Robustness tests further show that blur and center occlusion are the dominant failure modes, while HOG features remain valuable under illumination changes. Future work should repeat the same protocol on the official training archive, add learned visual backbones, and connect calibrated uncertainty to closed-loop driving decisions.

References

[1] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "The German Traffic Sign Recognition Benchmark: A multi-class classification competition," in Proc. IEEE Int. Joint Conf. Neural Networks, 2011, pp. 1453-1460.

[2] Jinyi Mu, Yifei Lu, and Michelle Smith, “LLM-Assisted Incrementality (Uplift) Modeling for Email Advertising: From Feature Interactions to Interpretable Audience–Creative–Channel Policies ”, JACS, vol. 3, no. 1, pp. 31–48, Jan. 2023, doi: 10.69987/JACS.2023.30103.

[3] S. Houben, J. Stallkamp, J. Salmen, M. Schlipsing, and C. Igel, "Detection of traffic signs in real-world images: The German Traffic Sign Detection Benchmark," in Proc. Int. Joint Conf. Neural Networks, 2013, pp. 1-8.

[4] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2005, pp. 886-893.

[5] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proc. IEEE, vol. 86, no. 11, pp. 2278-2324, 1998.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Proc. Advances in Neural Information Processing Systems, 2012, pp. 1097-1105.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. Int. Conf. Learning Representations, 2015.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2016, pp. 770-778.

[9] C. Szegedy et al., "Going deeper with convolutions," in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2015, pp. 1-9.

[10] Yuanzheng Chen, Yitian Zhang, David Chau, and Matt Sherman, "Credit Card Default Risk Tiering with Probability Calibration and Uncertainty-Driven Rejection: A Reproducible Study on the UCI Credit Card Clients Dataset", JACS, vol. 3, no. 4, pp. 31–47, Apr. 2023, doi: 10.69987/JACS.2023.30403.

[11] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. Int. Conf. Machine Learning, 2019, pp. 6105-6114.

[12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in Proc. IEEE Int. Conf. Computer Vision, 2017, pp. 2980-2988.

[13] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," J. Machine Learning Research, vol. 15, pp. 1929-1958, 2014.

[14] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in Proc. Int. Conf. Machine Learning, 2015, pp. 448-456.

[15] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. Int. Conf. Machine Learning, 2017, pp. 1321-1330.

[16] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in Proc. Int. Conf. Machine Learning, 2005, pp. 625-632.

[17] A. Ovadia et al., "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift," in Proc. Advances in Neural Information Processing Systems, 2019, pp. 13991-14002.

[18] Siming Zhao, Hailin Zhou, and Daniel Martinez, "LLM-Assisted Causal Attribution of Service Performance Upgrades on Churn and Tenure: Full Evaluation on the IBM Telco Customer Churn Dataset", JACS, vol. 3, no. 2, pp. 18–34, Feb. 2023, doi: 10.69987/JACS.2023.30202.

[19] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in Proc. Int. Conf. Machine Learning, 2016, pp. 1050-1059.

[20] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in Proc. Advances in Neural Information Processing Systems, 2017, pp. 5574-5584.

[21] A. Radford et al., "Learning transferable visual models from natural language supervision," in Proc. Int. Conf. Machine Learning, 2021, pp. 8748-8763.

[22] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "VisualBERT: A simple and performant baseline for vision and language," arXiv:1908.03557, 2019.

[23] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in Proc. Advances in Neural Information Processing Systems, 2019, pp. 13-23.

[24] H. Tan and M. Bansal, "LXMERT: Learning cross-modality encoder representations from transformers," in Proc. Conf. Empirical Methods in Natural Language Processing, 2019, pp. 5100-5111.

[25] X. Li et al., "Oscar: Object-semantics aligned pre-training for vision-language tasks," in Proc. European Conf. Computer Vision, 2020, pp. 121-137.

[26] W. Kim, B. Son, and I. Kim, "ViLT: Vision-and-language transformer without convolution or region

supervision," in Proc. Int. Conf. Machine Learning, 2021, pp. 5583-5594.

[27] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in Advances in Large Margin Classifiers, 1999, pp. 61-74.

[28] Daren Zheng, Chenyu Li, and Harvey Davidson, “Continual Red-Teaming for In-the-Wild Jailbreaks via Online Guardrail Updates and Guardrail Distillation”, JACS, vol. 3, no. 2, pp. 35–49, Feb. 2023, doi: 10.69987/JACS.2023.30203.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. Int. Conf. Learning Representations, 2015.

[30] G. Bradski, "The OpenCV Library," Dr. Dobb’s Journal of Software Tools, vol. 25, no. 11, pp. 120-125, 2000.