ICCV 2025
DriveX
#XXXX

Understanding What Vision-Language Models See in Traffic: PixelSHAP for Object-Level Attribution in Autonomous Driving

Anonymous ICCV 2025 DriveX submission

Paper ID XXXX

Abstract

Vision-Language Models (VLMs) are increasingly used in autonomous driving for scene understanding, hazard detection, and decision-making support. Yet, knowing which traffic objects these models prioritize is crucial for safety validation and trust. Existing interpretability methods provide pixel-level attributions but fail to answer the key question: "Which specific objects—vehicles, pedestrians, traffic signs—influence the model's driving decisions?"

*We introduce **PixelSHAP**, a model-agnostic framework for object-level explainability in Vision-Language Models applied to traffic scenarios. PixelSHAP extends Shapley-based attribution to structured visual entities, systematically quantifying how individual traffic participants influence a VLM's reasoning about driving situations. Operating purely on input-output behavior, our method is compatible with both open-source models (LLaVA, LLaMA-Vision) and commercial systems (GPT-4V, Gemini) commonly used in autonomous driving applications.*

Our approach introduces novel masking strategies including Bounding Box with Overlap Avoidance (BBOA) that address fundamental challenges in traffic scene attribution, achieving complete object occlusion while minimizing interference with neighboring vehicles or infrastructure. We evaluate PixelSHAP on traffic scene understanding tasks, demonstrating its ability to reveal which objects VLMs prioritize for different driving scenarios. Compared to simple baselines, PixelSHAP provides semantically meaningful attributions that align with human expectations about traffic safety priorities.

Beyond technical contribution, PixelSHAP enables safety engineers to audit VLM behavior in autonomous driving contexts, identify potential failure modes, and validate that models focus on safety-critical objects. Our implementation provides immediate practical value for developing more transparent and trustworthy autonomous driving systems.

1. Introduction

Vision-Language Models (VLMs) are increasingly integral to autonomous driving systems, supporting scene understanding, hazard detection, and driving decision assistance. As these models move from prototypes to safety-critical deployments, understanding their decision-making is crucial for ensuring passenger safety and public trust.

Consider a scenario: a VLM analyzes a busy intersection and outputs: "Pedestrian visible in crosswalk, vehicle should yield." This triggers braking protocols. Yet, among multiple pedestrians—on sidewalks, near the crosswalk, and one crossing—which person influenced the decision? Attribution is essential to validate that the system responded to the correct participant.

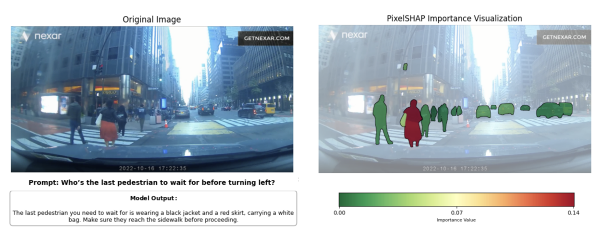

Figure 1. PixelSHAP reveals object-level attribution in traffic scenes. It identifies which pedestrian influenced the VLM's safety assessment, enabling validation that the model focused on the actual crossing pedestrian.

The core challenge is the semantic gap between how VLMs process visual information and how we interpret their decisions for safety validation. Existing methods fall short: gradient-based approaches like GradCAM require model internals unavailable in commercial VLMs, while pixel-level perturbation methods such as RISE blur distinct traffic participants into indecipherable importance regions.

We propose PixelSHAP, a model-agnostic framework for object-level interpretability in traffic scenarios. Extending Shapley value attribution from tokens to structured visual entities, PixelSHAP quantifies how vehicles, pedestrians, traffic signs, and infrastructure influence VLM assessments.

A key innovation is our object-level perturbation approach. Our Bounding Box with Overlap Avoidance (BBOA) achieves complete object occlusion while preserving context for neighboring elements, enabling clean attribution critical for safety validation.

This paper makes four contributions to interpretable AI for autonomous driving:

- **Traffic-Focused Object-Level Attribution:** A framework identifying traffic participants influencing VLM decisions, compatible with open-source and commercial models.
- **BBOA Masking Strategy:** A perturbation method that occludes target objects while preserving surrounding context.
- **Multi-Model Validation:** Evaluation across four VLMs, comparing against adapted interpretability baselines.
- **Traffic Scene Evaluation:** Protocols for assessing attribution quality in driving-relevant contexts.

The remainder of this paper describes our methodology, experimental validation, and implications for interpretable autonomous driving systems.

2. Related Work

Understanding VLM decisions in traffic scenarios requires explainability methods that can identify which specific visual objects influence model outputs. The choice of explainability approach is fundamentally constrained by model accessibility and the semantic granularity required for safety validation in autonomous driving applications.

2.1. White-Box vs. Black-Box Explainability

Explainability methods for VLMs divide into white-box approaches requiring access to model internals and black-box methods operating solely on input-output behavior. White-box methods like Grad-CAM [14] analyze internal gradients and activations to generate attribution maps. LVLM-Interpret [16] provides attention visualization, relevancy maps, and causal interpretation for vision-language models by accessing transformer weights and gradients.

White-box methods offer detailed insights into model mechanisms but face limitations for practical autonomous driving applications. Many state-of-the-art VLMs deployed in commercial autonomous systems, including GPT-4V [7] and Gemini-2.0 [1], do not provide access to internal weights or gradients. For applications requiring interpretability of production-deployed models, black-box approaches become essential.

2.2. Black-Box Perturbation-Based Methods

Black-box methods explain model decisions through systematic input perturbation and output analysis, making them compatible with any VLM regardless of architecture. RISE [10] generates importance maps by randomly masking image regions and measuring output changes. LIME [13] learns local linear approximations around input instances using perturbation-based sampling.

These pixel-level approaches face limitations when analyzing traffic scenes with multiple objects. When vehicles, pedestrians, and infrastructure appear in proximity, pixel-based attribution creates blended importance maps that cannot isolate individual traffic participants. For autonomous driving safety validation, understanding which specific object influenced a model's assessment requires object-level granularity that pixel-based methods cannot provide.

2.3. Shapley Values for Principled Attribution

Shapley values from cooperative game theory [15] provide mathematically principled feature attribution with desirable properties including efficiency, symmetry, and additivity. TokenSHAP [3] demonstrated their effectiveness for language model interpretability by quantifying individual token contributions. MM-SHAP [8] applied Shapley values to multimodal models, measuring the relative importance of visual versus textual modalities using image patches.

While Shapley-based approaches offer theoretical rigor, existing applications focus on different granularities and questions than object-level attribution in traffic scenarios. Extending Shapley principles to semantic object-level analysis while maintaining black-box compatibility remains an active area of development.

2.4. Multimodal Interpretability: Related Approaches and Distinctions

Recent interpretability frameworks for VLMs address complementary aspects of multimodal understanding, though with different focus areas than object-level attribution:

MM-SHAP [8] provides valuable insights into modality-level contributions, quantifying whether models rely more on textual or visual information. However, its patch-based granularity cannot isolate individual traffic participants within scenes. While MM-SHAP can reveal that a model used "60% vision, 40% text," it cannot distinguish which specific vehicle or pedestrian drove that visual contribution—a distinction critical for autonomous driving safety validation.

LVLM-Interpret [16] offers comprehensive analysis through attention visualization and causal interpretation, providing detailed insights into model reasoning processes. However, its dependency on white-box access to attention weights and gradients limits applicability to commercial VLMs commonly deployed in autonomous systems. Additionally, its patch-based visualizations operate at spatial resolutions that may not align with semantic object boundaries essential for traffic safety analysis.

ICCV 2025 DriveX #XXXX

These methods address important questions about multimodal reasoning and provide valuable debugging capabilities. Our work complements these approaches by focusing specifically on the object-level attribution question that existing methods cannot directly address due to granularity and accessibility constraints.

2.5. Object-Level Attribution: Addressing the Granularity Gap

Current black-box methods cannot directly answer questions critical for traffic scene understanding: "Which specific vehicle influenced the model's safety assessment?" or "Did the model focus on the crossing pedestrian or background elements?" This limitation stems from the granularity mismatch between available attribution methods (pixels or patches) and the semantic units relevant for autonomous driving validation (objects representing traffic participants).

The autonomous driving context amplifies these challenges because safety validation requires understanding attribution at the semantic level of traffic participants—vehicles, pedestrians, cyclists, and infrastructure—rather than abstract visual regions. Existing pixel-level methods cannot distinguish between a pedestrian actively crossing versus one standing on a sidewalk when both appear in the same image region, yet this distinction is critical for validating autonomous driving decisions.

2.6. Our Approach

We introduce PixelSHAP to address the object-level attribution gap by extending Shapley-based attribution to individual traffic objects while maintaining black-box compatibility with commercial VLMs. Our approach builds on the theoretical foundation of Shapley values while adapting the methodology to operate on semantic objects rather than pixels or patches.

PixelSHAP complements existing interpretability methods by focusing on the specific granularity and accessibility requirements of autonomous driving applications. We demonstrate improvements over adapted versions of existing methods (RISE-Objects) and simple heuristics, showing that principled Shapley attribution can provide more accurate object-level explanations for traffic safety validation. Our evaluation includes comparison with gradient-based methods where applicable, providing insight into the relative performance of black-box versus white-box approaches for object-level attribution tasks.

3. Problem Statement

We formalize object-level attribution in Vision-Language Models (VLMs) as a black-box interpretability challenge: quantifying how individual visual objects contribute to a model's textual output.

3.1. Problem Formulation

Given a VLM f mapping an image I and optional text prompt p to a response $y = f(I, p)$, our goal is to assign an attribution score ϕ_i to each object o_i in $O = \{o_1, o_2, ..., o_n\}$, representing its influence on y. Attribution scores must satisfy:

1. **Efficiency**: $\sum_{i=1}^{n} \phi_i = f(I, p) - f(\emptyset, p)$, where $\emptyset$ is the scene with all objects removed.
2. **Symmetry**: Identical contributors receive equal scores.
3. **Additivity**: Scores combine consistently across object subsets.

3.2. Key Constraints and Requirements

Black-Box Compatibility: The method must function without access to model internals, gradients, or attention weights, ensuring compatibility with commercial VLMs. **Object-Level Granularity**: Beyond pixel-level maps, we require semantic object attribution to answer, e.g., "Which specific vehicle influenced the decision?" **Semantic Preservation**: Perturbations must fully remove an object's contribution while maintaining scene context.

3.3. Applications and Use Cases

This formulation supports critical interpretability needs: In *autonomous systems*, identifying which traffic participants (vehicles, pedestrians, signs) influenced a VLM's assessment validates correct prioritization of safety-critical objects. In *content moderation*, it clarifies which visual elements trigger policy violations, improving automated systems. In *medical imaging*, object-level attribution aids in validating diagnostic outputs and building clinician trust. In *general scene understanding*, it verifies that VLMs attend to relevant elements rather than spurious correlations.

3.4. Technical Challenges

Object Segmentation Dependency: Reliable attribution depends on accurate detection and segmentation of objects. **Occlusion Strategy**: Removing an object cleanly while preserving scene context requires sophisticated masking to avoid artifacts or distortion of neighboring elements. **Computational Efficiency**: Exact Shapley value computation is infeasible; efficient approximations are essential. **Evaluation Methodology**: Assessing attribution quality requires ground truth aligned with human judgments of object importance.

The following sections describe how PixelSHAP addresses these challenges.

4. Methodology

PixelSHAP extends Shapley value attribution from textual tokens to visual objects, enabling principled object-level interpretability for Vision-Language Models. Our approach operates through three stages: object identification with segmentation, systematic perturbation, and attribution computation.

4.1. Framework Design

The framework requires both object detection (bounding boxes) and segmentation masks for each object. Users can integrate results from any detection system suited to their application domain, including category-specific models like YOLO [11] variants or open-vocabulary systems like GroundingDINO [6]. When detection systems provide only bounding boxes, we automatically generate segmentation masks using SAM2 [4] within the detected regions to ensure complete object-level analysis.

We formulate object attribution as a cooperative game where detected objects serve as players and the VLM's response represents the outcome. For objects $O = \{o_1, o_2, ..., o_n\}$, each object's Shapley value ϕ_i quantifies its contribution:

$$\phi_i = \sum_{S \subseteq O \setminus \{o_i\}} \frac{|S|!(|O| - |S| - 1)!}{|O|!} [v(S \cup \{o_i\}) - v(S)]$$

where $v(S)$ measures the VLM's response when only objects in subset S remain visible.

4.2. Object Perturbation Strategy

The central challenge lies in removing target objects while preserving scene context for accurate attribution. We propose Bounding Box with Overlap Avoidance (BBOA) and evaluate it against two established baselines.

Precise masking applies exact segmentation boundaries but creates irregular occlusions that may introduce visual artifacts. Bounding box masking uses rectangular regions but risks occluding adjacent objects in dense scenes.

BBOA combines the advantages of both approaches through a three-step process: first masking the target object's bounding box region, then identifying other objects whose segmentation masks intersect this region, and finally restoring those overlapping objects by unmasking their precise boundaries. This strategy ensures complete target removal while preserving neighboring objects regardless of scene density.

4.3. Computational Implementation

Exact Shapley computation requires evaluating 2^n object subsets, which becomes computationally prohibitive for scenes with many objects. We employ sampling-based approximation that reduces VLM queries from exponential to linear scaling, typically requiring 100-300 evaluations for scenes with 10-15 objects and completing analysis within 30-60 seconds.

Response similarity is measured using semantic embedding approaches through sentence transformers [12] or lexical similarity metrics depending on application requirements. The framework operates entirely through VLM input-output interfaces, maintaining compatibility with both open-source and commercial models without requiring access to internal representations.

5. Experimental Evaluation

We evaluate PixelSHAP's effectiveness for object-level attribution in vision-language models through systematic comparison with existing black-box methods on carefully constructed human-annotated datasets.

5.1. Dataset Construction

The absence of suitable benchmarks for object-level VLM attribution necessitated creating specialized evaluation datasets. We developed two complementary datasets with human annotation protocols designed to assess object-level interpretability across different visual domains.

BDD10K Traffic Dataset: We selected 250 representative images from the Berkeley DeepDrive dataset (BDD10K) [18], focusing on driving scenarios containing multiple traffic participants (vehicles, pedestrians, cyclists, traffic signs). Each scene was chosen to represent common driving situations where understanding object-level attention becomes safety-critical: intersections with multiple vehicles, crosswalks with pedestrians, and complex urban environments with mixed traffic.

Three experienced annotators followed a structured protocol: first, they randomly selected one object from each traffic scene, then formulated driving-relevant questions that would require focusing on that specific object to answer correctly (e.g., "Which vehicle poses the greatest safety concern?" or "What traffic element should influence the driver's next action?"). This approach ensures unbiased ground truth while maintaining realistic question formulation. We measured inter-annotator agreement using Fleiss' kappa [2] and retained only scenes achieving substantial consensus ($\kappa > 0.7$). Figure 3 illustrates representative examples from this dataset, showing the diversity of objects and question types used in evaluation.

COCO General Dataset: To demonstrate broader applicability beyond traffic scenarios, we created a complementary dataset using 250 images selected from COCO [5] validation set. Following identical annotation protocols, annotators first randomly selected objects from general visual scenes, then generated focused questions requiring attention to those specific objects. This dataset enables assessment of PixelSHAP's effectiveness across diverse visual contexts while maintaining the same evaluation framework.

Both datasets are publicly available on Hugging

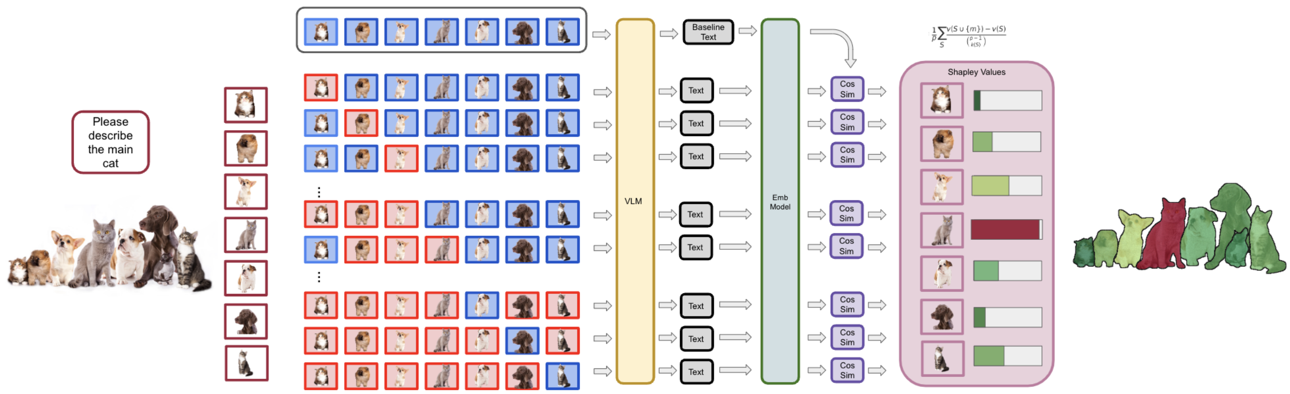

Figure 2. Overview of the PixelSHAP framework. The method systematically perturbs object groups, queries a vision-language model (VLM), and computes Shapley values to quantify object importance.

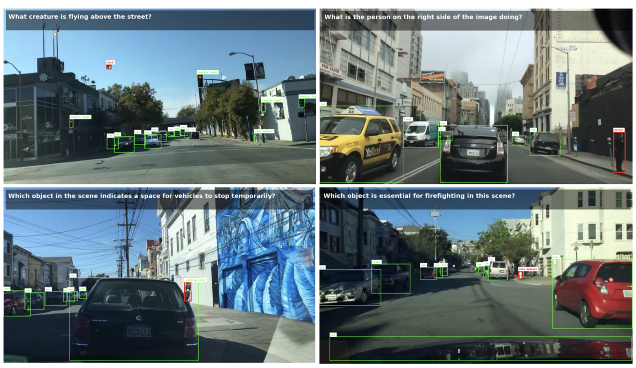

Figure 3. Sample annotations from BDD10K Traffic Dataset showing diverse object types and corresponding questions. Each example demonstrates how human annotators formulated questions requiring attention to specific objects for accurate answering.

Face[17], providing standardized benchmarks for future research in object-level VLM interpretability.

5.2. Evaluation Protocol

Our evaluation protocol measures how well attribution methods identify the same objects that human experts consider most relevant for answering given questions. For each image-question pair, we provide the question and corresponding answer to the VLM, then apply different attribution methods to identify which objects the model should focus on. We compare these attributions against human annotations to assess attribution quality.

This framework enables direct comparison of different attribution approaches while maintaining consistency with human reasoning patterns about object relevance in visual question answering tasks.

5.3. Masking Strategies

PixelSHAP's effectiveness depends critically on the masking strategy used during object occlusion. We investigate two primary approaches for handling object removal during attribution computation:

Precise Masking: Objects are masked exactly according to their segmentation boundaries, replacing object pixels with neutral background or inpainting [9]. This approach maintains precise object boundaries but may create artificial visual artifacts at object edges.

Bounding Box Occlusion with Adjustment (BBOA): Objects are occluded using expanded bounding boxes that fully contain the object while minimizing overlap with neighboring objects. This strategy avoids edge artifacts and prevents unintended masking of adjacent objects that might confound attribution computation.

Figure 4 illustrates these different masking approaches and their impact on attribution quality. The BBOA strategy demonstrates superior performance by ensuring complete object occlusion while preserving the integrity of surrounding visual context.

5.4. Baseline Comparison Framework

We establish PixelSHAP's effectiveness through comparison with existing black-box interpretability methods adapted for object-level analysis. Since direct comparison requires operating at the same semantic granularity, we adapt pixel-level methods to produce object-level attributions.

RISE-Objects: The original RISE method [10] generates pixel-level importance maps through random masking. We adapt RISE to operate at object-level granularity by randomly masking subsets of detected objects and measuring resulting changes in model output similarity. This preserves RISE's core perturbation methodology while enabling fair

ICCV 2025 DriveX #XXXX

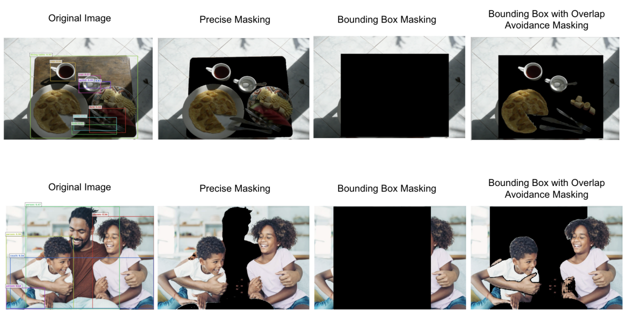

Figure 4. Comparison of masking strategies for object occlusion in PixelSHAP. (a) Precise masking follows exact segmentation boundaries. (b) Bounding Box Occlusion with Adjustment (BBOA) uses expanded boxes to ensure complete occlusion while minimizing interference with neighboring objects. BBOA consistently achieves better attribution performance across different scenarios.

comparison at the semantic object level.

Simple Heuristic Baselines: We include largest object (by bounding box area) and central object (closest to image center) as basic attribution methods. These baselines test whether sophisticated attribution approaches provide meaningful improvements over simple assumptions about visual attention.

Random Baseline: Random object selection establishes the performance floor and validates that our evaluation metrics capture meaningful attribution quality differences.

5.5. Evaluation Metrics

We assess attribution quality using metrics aligned with human annotation protocols and practical interpretability needs:

Recall@1: Percentage of test cases where the highest-attributed object matches human expert annotation. This metric directly measures whether attribution methods identify the same object that human experts consider most relevant.

Recall@3: Percentage where the human-annotated target object appears among the top-3 attributed objects, providing insight into attribution ranking quality.

Mean Reciprocal Rank (MRR): Average inverse rank of the ground-truth object across all test cases, offering a nuanced view of attribution accuracy that accounts for ranking position.

5.6. Results and Analysis

Table 1 presents comprehensive performance comparison across four representative VLMs on both datasets. The results demonstrate that PixelSHAP with BBOA masking achieves the best performance in nearly all scenarios, though some competitive cases reveal interesting model-specific characteristics.

Model-Specific Performance Patterns: Gemini-2.0-flash achieves the highest overall performance across both datasets, with particularly strong results on COCO general scenes (67.48% Recall@1) and leading performance on traffic scenarios (64.7% Recall@1). GPT-4o demonstrates competitive performance on traffic scenarios (63.8% Recall@1), while both LLaVA-v1.5-7B and LLaMA-3.2-11B-Vision show more modest but consistent results across datasets.

Masking Strategy Analysis: BBOA achieves the best performance in the vast majority of cases, though some notable exceptions highlight the complexity of optimal masking strategies. LLaMA-3.2-11B-Vision shows a rare case where precise masking slightly outperforms BBOA on traffic Recall@1 (56.1% vs 55.8%), while LLaVA-v1.5-7B demonstrates competitive performance where precise masking achieves higher Recall@3 and MRR scores on traffic scenarios. These close margins suggest that masking strategy optimization may be model-dependent in specific contexts.

Attribution Method Robustness: The consistently strong performance of BBOA across different models and datasets validates our approach, with typical improvements of 3-8 percentage points over precise masking and 10-20 percentage points over baseline methods. The few competitive cases where precise masking approaches BBOA performance (difference 0.3 percentage points) demonstrate that while BBOA is generally superior, the optimal strategy may require fine-tuning for specific model architectures.

Baseline Comparison: PixelSHAP variants substantially outperform simple heuristics and RISE-Objects across all conditions. RISE-Objects achieves moderate performance but consistently lags behind PixelSHAP approaches by 5-15 percentage points in Recall@1, confirming the benefits of principled Shapley-based attribution for object-level interpretability.

Dataset-Specific Insights: Performance patterns show interesting domain dependencies. Gemini-2.0-flash maintains strong advantages on both datasets, suggesting robust generalization capabilities. The traffic scenarios generally yield slightly higher absolute performance across models, potentially reflecting the more structured nature of driving scenes compared to diverse COCO imagery.

5.7. Computational Efficiency

PixelSHAP processing requires 35-65 seconds per image depending on object count and VLM inference speed, using approximately 2-3× the number of detected objects in API calls rather than the exponential scaling that naive Shapley computation would require. This represents practical computational requirements suitable for offline analysis and

Model	Method	BDD10K Traffic Dataset			COCO General Dataset		
		R@1 (%)	R@3 (%)	MRR	R@1 (%)	R@3 (%)	MRR
GPT-4o	PixelSHAP (BBOA)	**63.8**	**86.2**	**0.75**	**60.56**	**87.66**	**0.73**
	PixelSHAP (Precise)	59.2	82.1	0.71	57.61	84.71	0.69
	PixelSHAP (BBox)	55.7	78.4	0.67	53.18	85.20	0.68
	RISE-Objects	43.1	67.8	0.57	42.3	69.2	0.56
Gemini-2.0-flash	PixelSHAP (BBOA)	**64.7**	**85.9**	**0.76**	**67.48**	**89.17**	**0.77**
	PixelSHAP (Precise)	62.1	83.2	0.73	59.62	84.73	0.71
	PixelSHAP (BBox)	59.4	80.8	0.71	58.10	88.68	0.72
	RISE-Objects	47.6	72.1	0.61	45.7	71.6	0.59
LLaVA-v1.5-7B	PixelSHAP (BBOA)	**48.9**	71.4	0.61	**49.78**	**83.28**	**0.65**
	PixelSHAP (Precise)	48.2	**72.1**	**0.62**	49.27	75.38	0.61
	PixelSHAP (BBox)	45.6	68.9	0.59	43.88	76.32	0.59
	RISE-Objects	41.3	65.7	0.55	37.2	64.5	0.52
LLaMA-3.2-11B-Vision	PixelSHAP (BBOA)	55.8	**78.3**	**0.68**	**52.71**	**86.72**	**0.68**
	PixelSHAP (Precise)	**56.1**	77.9	0.68	49.76	80.27	0.64
	PixelSHAP (BBox)	53.4	76.2	0.66	50.76	80.32	0.65
	RISE-Objects	44.8	68.5	0.58	38.9	66.4	0.53
Largest Object		38.4	62.5	0.51	23.14	60.85	0.43
Central Object		31.7	58.1	0.46	36.92	70.62	0.52

Table 1. Object-level attribution performance comparison across VLMs and datasets. Results show mean performance over test sets. Bold indicates best performance for each model-method combination.

safety validation applications in autonomous driving systems.

5.8. Limitations and Future Work

Our evaluation reveals several limitations that inform future research directions. Performance degrades in extremely cluttered scenes (> 15 objects) where occlusion becomes pervasive. Attribution quality also depends on segmentation accuracy, creating dependency on upstream computer vision components.

Segmentation Quality Impact: We evaluate attribution degradation under noisy segmentation by introducing controlled errors (10-30% mask boundary deviation) to ground-truth objects. Performance drops 8-15% with moderate noise, confirming segmentation dependency while demonstrating reasonable robustness to realistic segmentation errors.

The varying performance patterns across models suggest that future work should explore model-specific attribution strategies that account for architectural differences in visual reasoning capabilities.

5.9. Qualitative Examples

Figure 5 demonstrates PixelSHAP's ability to provide intuitive, context-sensitive attributions across different query types and scenarios.

Figure 5. PixelSHAP attribution examples across traffic and general scenarios. Each row shows the same scene analyzed with different questions, demonstrating context-sensitive attribution. Red intensity indicates object importance scores.

Traffic Scene Analysis: In driving scenarios, PixelSHAP correctly prioritizes safety-critical objects based on query context. When asked "Which vehicle should the driver monitor?", the method emphasizes the approaching car rather than parked vehicles. For pedestrian-focused queries like "Is it safe to proceed?", attribution shifts to highlight the person near the crosswalk while de-emphasizing background traffic.

Context Sensitivity: The examples demonstrate sophisticated adaptation to query specificity. Identical visual scenes produce different attribution patterns when analyzed with different questions. General queries about scene content distribute attention across multiple objects, while specific queries about particular object types concentrate attribution on relevant entities.

General Scene Understanding: Beyond traffic applications, PixelSHAP provides meaningful attributions for diverse visual reasoning tasks. When analyzing animal scenes, queries about "What animals are present?" appropriately emphasize biological entities while ignoring background objects. Action-focused questions shift attribution toward objects involved in activities rather than static scene elements.

These qualitative results confirm that PixelSHAP captures the contextual reasoning that makes VLM interpretability valuable for practical applications. The method's ability to adapt attribution patterns based on query intent enables users to understand not just what objects are visually prominent, but which objects actually influence the model's reasoning for specific tasks.

6. Discussion

Our work fills a critical gap in understanding how Vision-Language Models (VLMs) reason about traffic scenes, offering practical tools for safety validation in autonomous driving.

6.1. Key Findings

PixelSHAP extends Shapley-based attribution to object-level analysis while maintaining black-box compatibility, essential for commercial VLMs. The BBOA masking strategy resolves the core challenge of occluding target objects without interfering with surrounding context, outperforming existing perturbation methods.

Consistent performance across diverse VLM architectures indicates that our approach captures fundamental aspects of vision-language reasoning for traffic understanding. While commercial models achieve higher absolute attribution accuracy, relative gains from object-level analysis remain comparable across architectures.

For autonomous driving, PixelSHAP allows engineers to verify that VLMs attend to the correct traffic participants, supporting validation workflows critical for safe deployment.

6.2. Limitations

Attribution accuracy depends on segmentation quality, making it reliant on upstream detection performance. Current processing times (35–65 seconds per image) suit offline analysis but limit real-time use.

Though BBOA minimizes distribution shift, masking can still alter image statistics, especially in cluttered scenes with over 15 objects. Evaluation relies on human annotations, which may not fully capture expert safety priorities.

6.3. Future Directions

Integration with Autonomous Systems: Embedding attribution into development workflows could enable continuous validation during system updates. **Temporal Analysis:** Extending to video sequences would support reasoning about dynamic object importance. **Domain-Specific Models:** Adapting the framework for traffic-specific categories (e.g., emergency vehicles, construction zones) could improve safety-critical assessments. **Computational Optimization:** Approximation models may accelerate attribution, enabling iterative development workflows.

7. Conclusion

We introduced PixelSHAP, a model-agnostic framework for object-level attribution in Vision-Language Models for traffic scene understanding. By extending Shapley value attribution to structured visual entities, our approach provides meaningful explanations that enable safety validation in autonomous driving.

Through systematic evaluation, we demonstrated consistent improvements over prior methods. Our BBOA masking strategy addresses key challenges in perturbation-based attribution while supporting compatibility with commercial VLMs.

PixelSHAP empowers engineers to identify which traffic participants influence model decisions, supporting safety validation workflows for autonomous systems. Our open-source implementation facilitates adoption in research and practice. Future work should focus on computational efficiency and integration into autonomous vehicle development pipelines to enhance transportation safety.

Supplementary Material

Additional materials including code, datasets, and extended results will be made available upon acceptance to maintain anonymity during the review process.

References

[1] Google DeepMind. Gemini 2.0 pro, 2024. Released as part of Gemini AI Suite. 2

[2] Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971. 4

[3] Miriam Horovicz and Roni Goldshmidt. Tokenshap: Interpreting large language models with monte carlo shapley value estimation. In *EMNLP NLP4Science Workshop*, 2024. 2

[4] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Yuxin Rolland, Linus Gustafson, Trevor Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv preprint*, 2023. 4

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. *arXiv preprint*, abs/1405.0312, 2014. 4

[6] Shilong Liu, Feng Li, Hao Zhang, Xiao Zhang, Lei Zhu, Hang Wang, Jianlong Shi, Hongyang Li, and Hao Dong. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint*, 2023. 4

[7] OpenAI. Gpt-4o model card, 2024. Accessed via OpenAI API. 2

[8] L. Parcalabescu and A. Frank. Mm-shap: Multimodal shapley values for model interpretation. *arXiv preprint*, 2022. 2

[9] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2536–2544, 2016. 5

[10] V. Petsiuk, A. Das, and K. Saenko. Rise: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference (BMVC)*, 2018. 2, 5

[11] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016. 4

[12] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019. 4

[13] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you? explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, page 1135–1144, 2016. 2

[14] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Int. Conf. Comput. Vis.*, pages 618–626, 2017. 2

[15] Lloyd S. Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2:307–317, 1953. 2

[16] G. B. M. Stan, R. Saunshi, N. Viswanathan, G. Sundaramoorthi, and A. Menon. Lvlm-interpret: An interpretability tool for large vision-language models. In *CVPR Workshop on Explainable AI for Computer Vision*, 2024. 2

[17] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Perric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020. 5

[18] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4