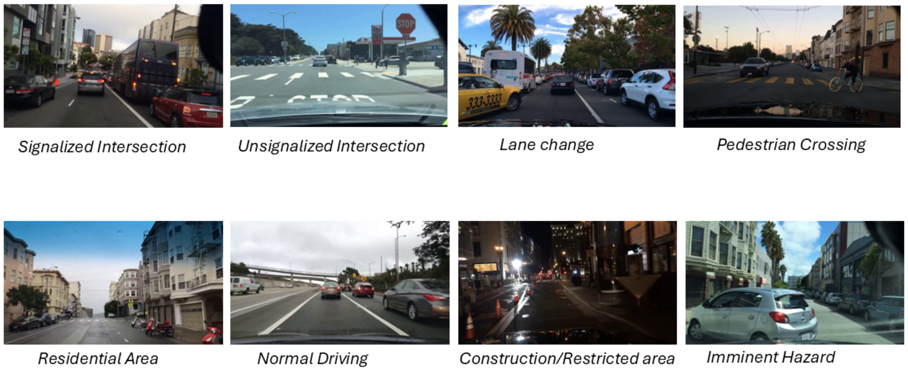


Signalized Intersection *Unsignalized Intersection* *Lane change* *Pedestrian Crossing*

Residential Area *Normal Driving* *Construction/Restricted area* *Imminent Hazard*

Fig. 7: Few-shot training scenarios. Our curated dataset covers eight critical driving contexts: signalized intersections, unsignalized intersections, lane changes, pedestrian crossings, residential areas, normal highway driving, construction zones, and imminent hazards to capture distinct gaze patterns essential for safe navigation [30, 49].

A Few-Shot Data Curation Strategy

A.1 Overview

We present a principled approach to curating few-shot training data that capture the temporal dynamics of driver attention. We develop a few-shot modeling framework that trains a general-purpose vision–language model to predict and explain the temporal dynamics of driver attention across diverse scenarios. Our approach identifies critical gaze transitions and extracts frame pairs that encode meaningful attention shifts in safety-critical scenarios across diverse driving scenarios. In this paper, we instantiate the framework on the BDD-A dataset, which consists of dash-cam video clips centered on driver-initiated braking events. From BDD-A, we curate 90 high-quality training samples spanning eight types of safety-critical driving contexts to demonstrate how human drivers reallocate attention in such scenarios. (Few examples are given in Figure 7).

A.2 Temporal Gaze Dynamics Mining

Our curation pipeline applies an information-theoretic analysis to the temporal structure of gaze patterns. Specifically, we compute the Kullback–Leibler (KL) divergence [50] between consecutive gaze maps, using KL peaks to pin-point moments of maximal change in the attention distribution. We treat these high-divergence moments as critical attention shifts and select frame pairs around them as the most informative training examples.

Parameter Selection. Algorithm 1 operates with four parameters, each motivated by properties of the BDD-A data and the temporal structure of driver gaze. $K{=}2$ bounds each clip's contribution to at most two pairs, preventing any

Algorithm 1 Temporal Gaze Transition Mining

```
Require: Synchronized gaze video $G$ and RGB video $V$ with $T$ frames
Require: Parameters $K = 2$, $\mathit{\Delta}_{\min} = 3$, $\mathit{\Delta}_{\max} = 18$, $r = 25$
Ensure: Frame pairs that maximize gaze transitions
for $t = 1$ to $T$ do
    $h_t \leftarrow \text{hist}(G_t)$
end for
for $t = 2$ to $T$ do
    $s_t \leftarrow \text{KL}(h_t \,\|\, h_{t-1})$
end for
$\bar{s} \leftarrow \text{smooth}(s, 5)$
$P \leftarrow \{\, t \mid \bar{s}_t > \bar{s}_{t-1} \;\wedge\; \bar{s}_t > \bar{s}_{t+1} \,\}$
$A \leftarrow \text{NMS}(P, \bar{s}, r)$                                   ▷ Enforce spacing
$\mathcal{C} \leftarrow \emptyset$
for each anchor $t \in A$ do
    $\delta_{\max} \leftarrow \min(\mathit{\Delta}_{\max}, T - t)$
    if $\delta_{\max} \geq \mathit{\Delta}_{\min}$ then
        $\delta_{\text{opt}} \leftarrow \arg \max_{\delta \in [\mathit{\Delta}_{\min}, \delta_{\max}]} \text{KL}(h_{t+\delta} \,\|\, h_t)$
        $\mathcal{C} \leftarrow \mathcal{C} \cup \{(t,\, t + \delta_{\text{opt}},\, \text{KL}(h_{t+\delta_{\text{opt}}} \,\|\, h_t))\}$
    end if
end for
Sort $\mathcal{C}$ by KL score in descending order
Select top $K$ pairs with mutual spacing at least $r$ frames
Export $\{(V_t, G_t, V_{t+\delta}, G_{t+\delta})\}$ for the selected pairs
```

single video from dominating the training distribution while maintaining temporal diversity, following temporal sampling strategies from video action recognition [5]. $\mathit{\Delta}_{\min}$=3 and $\mathit{\Delta}_{\max}$=18 define the temporal search window at 30fps, corresponding to 100–600ms — the range within which a driver fixation transition is perceptually meaningful and attributable to a single attention shift. r=25 enforces minimum anchor spacing to guarantee independence between selected pairs within the same clip. Applied to all 1000 BDD-A clips, after filtering clips shorter than 50 frames following [5], the algorithm yielded 651 candidate pairs; the remaining clips were excluded due to insufficient gaze dynamics or failure to satisfy the $\mathit{\Delta}_{\min}$ constraint within the available temporal window.

A.3 Structured Annotation Protocol

To generate consistent and informative captions, we developed a structured annotation protocol (Figure 8) that decomposes each gaze transition into four components:

- **Scene Context:** Establishes the environmental layout and traffic elements visible in the frame, providing spatial grounding for attention prediction.
- **Current gaze:** Identifies the primary attention target at time t based solely on the gaze heatmap, ensuring that annotations are grounded in actual gaze data.

- **Next Gaze:** Predicts the shift of attention to time $t + \Delta$, capturing the temporal evolution of gaze patterns.
- **Safety Rationale (Why):** Explains the driving-relevant motivation behind the attention transition, linking gaze behavior to safe navigation principles.

Each component is constrained to 25 words, enforcing conciseness while maintaining descriptive clarity. This rigid structure ensures consistency and preciseness across annotations while capturing the causal relationships essential for few-shot learning. The constraints encourage the model to capture fine-grained gaze differences across scenarios while enforcing a fixed output length, which helps reduce length-induced data bias.

A.4 Quality Control and Refinement

Our annotation pipeline combines automated generation with human refinement:

1. Initial annotations are generated using GPT-4o with the structured prompt
2. Human experts verify spatial accuracy against gaze heatmaps
3. Safety rationales are validated for driving relevance
4. Final annotations undergo consistency checks across similar scenarios

A.5 Dataset Statistics and Distribution

Table 4: Few-shot dataset composition across driving scenarios

Scenario Type	**Frame Pairs**	**Percentage**
Signalized Intersection	14	15.6%
Unsignalized Intersection	12	13.3%
Lane Change	13	14.4%
Pedestrian Crossing	11	12.2%
Residential Area	10	11.1%
Normal Driving	12	13.3%
Construction Zone	9	10.0%
Imminent Hazard	9	10.0%
Total	**90**	**100%**

The 651 candidate pairs were categorized into eight scenario types (Figure 7), unevenly distributed across categories, reflecting the natural frequency of each scenario type in BDD-A. To mitigate potential bias from category imbalance while constructing a slimmer few-shot training set, we applied stratified sampling to the candidate pairs. Specifically, within each of the eight scenario categories, pairs were ranked by KL divergence score and the top 30 pairs from each category were retained, following prior few-shot studies in semantic segmentation and object detection [12, 23, 60]. We further reduce the candidate pool to 9–14 pairs per category to minimize redundancy while preserving diversity in weather, lighting and scenario characteristics. This process resulted in 90 frame pairs with balanced representation across all eight scenario categories (Table 4). Each pair is then annotated with a structured caption following the protocol described in Section A.3, yielding one training sample per pair — capturing a complete attention transition from initial fixation through gaze shift to new target acquisition — providing rich supervision for few-shot learning.

A.6 Reproducibility

The project repository is available at `https://github.com/fsdam-vlc/fsdam`.

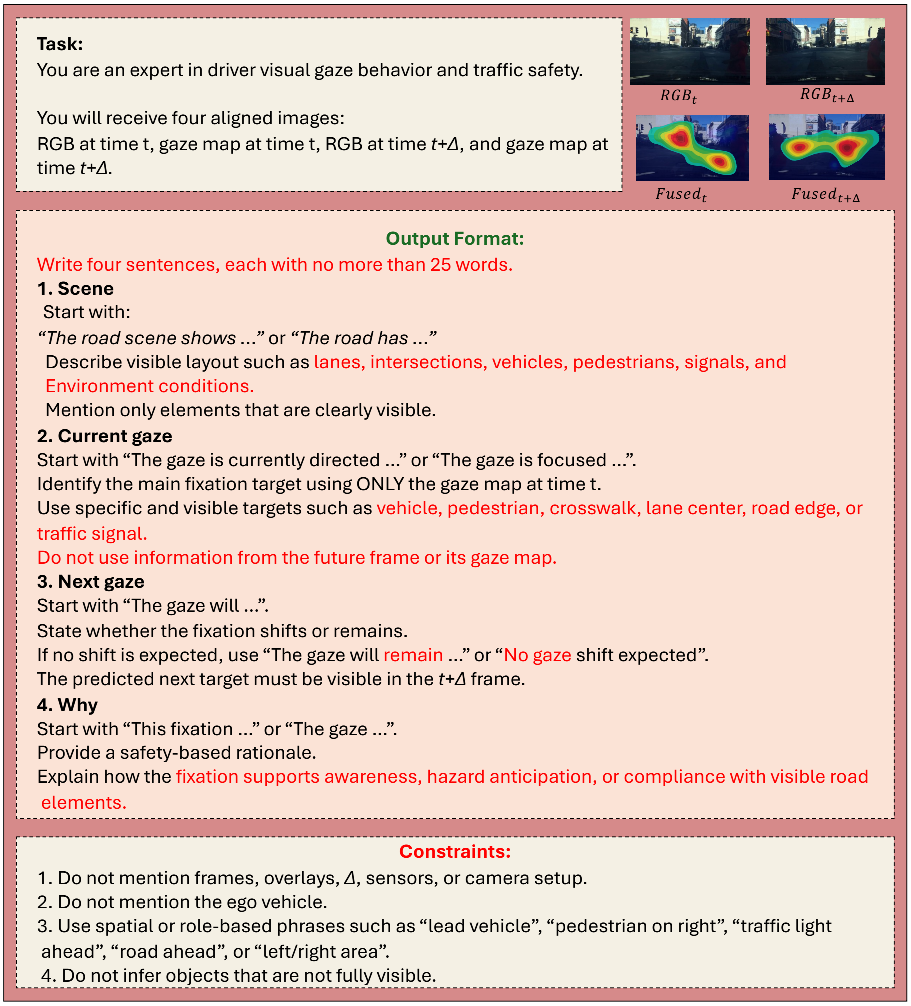
B Additional Qualitative Analysis

B.1 More examples on BDD-A Test Set

We provide comprehensive qualitative results demonstrating FSDAM's generalization on BDD-A test set, where ground truth captions are unavailable. Figure 9 shows representative examples across diverse urban scenarios.

FSDAM's gaze predictions closely match ground truth patterns across all examples, consistently focusing on the most immediate hazards. In the top row, both ground truth and prediction concentrate on the intersection ahead with approaching traffic. The middle example shows accurate attention on the lead vehicle while maintaining awareness of the pedestrian on the sidewalk. The bottom row demonstrates proper focus on the pedestrian crossing area at the intersection. The spatial distribution and intensity of predicted heatmaps align well with ground truth, validating our few-shot learning approach.

The generated captions remain semantically coherent, and even without corresponding ground-truth captions, their validity is reinforced by the qualitative consistency observed between caption content and gaze-attention heatmaps. Each follows our four-sentence structure: scene


RGB_t
$RGB_{t+\Delta}$
$Fused_t$
$Fused_{t+\Delta}$

Task:
You are an expert in driver visual gaze behavior and traffic safety.

You will receive four aligned images:
RGB at time t, gaze map at time t, RGB at time *t+Δ*, and gaze map at time *t+Δ*.

Output Format:

Write four sentences, each with no more than 25 words.

1. Scene
Start with:
"The road scene shows ..." or *"The road has ..."*
Describe visible layout such as lanes, intersections, vehicles, pedestrians, signals, and Environment conditions.
Mention only elements that are clearly visible.

2. Current gaze
Start with "The gaze is currently directed ..." or "The gaze is focused ...".
Identify the main fixation target using ONLY the gaze map at time t.
Use specific and visible targets such as vehicle, pedestrian, crosswalk, lane center, road edge, or traffic signal.
Do not use information from the future frame or its gaze map.

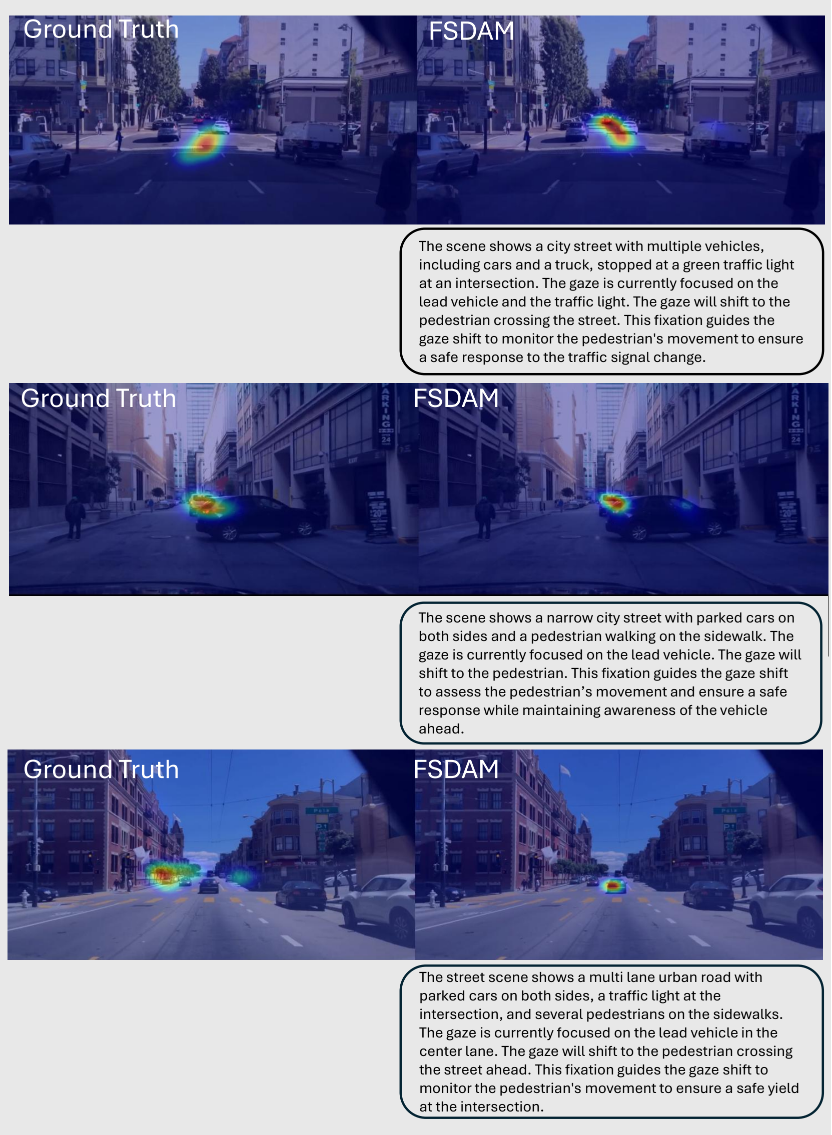
3. Next gaze
Start with "The gaze will ...".
State whether the fixation shifts or remains.
If no shift is expected, use "The gaze will remain ..." or "No gaze shift expected".
The predicted next target must be visible in the *t+Δ* frame.

4. Why
Start with "This fixation ..." or "The gaze ...".
Provide a safety-based rationale.
Explain how the fixation supports awareness, hazard anticipation, or compliance with visible road elements.

Constraints:

1. Do not mention frames, overlays, *Δ*, sensors, or camera setup.
2. Do not mention the ego vehicle.
3. Use spatial or role-based phrases such as "lead vehicle", "pedestrian on right", "traffic light ahead", "road ahead", or "left/right area".
4. Do not infer objects that are not fully visible.

Fig. 8: Structured annotation framework. Our four-sentence protocol ensures consistent, high-quality descriptions across all training pairs. Each sentence serves a specific purpose: (1) scene context establishment, (2) current attention state, (3) future attention prediction, and (4) safety-driven rationale. This structured approach enables the model to learn the causal relationship between scene dynamics and attention shifts.

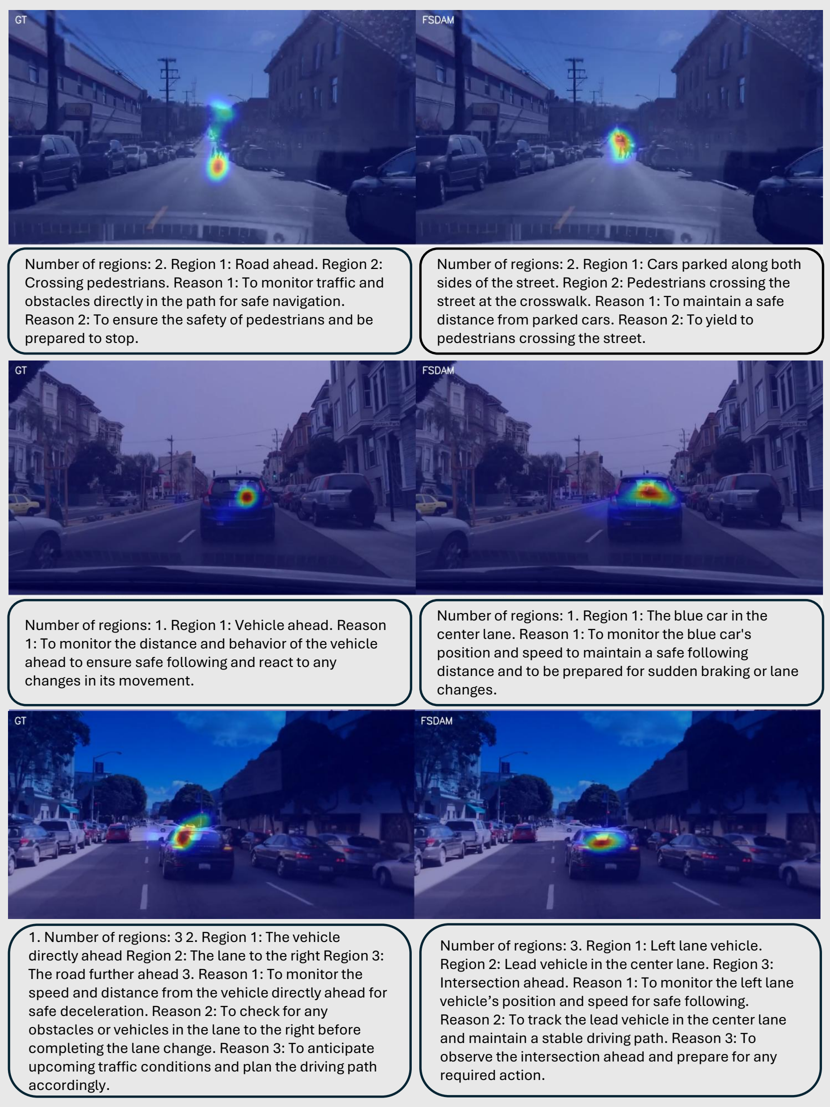


The scene shows a city street with multiple vehicles, including cars and a truck, stopped at a green traffic light at an intersection. The gaze is currently focused on the lead vehicle and the traffic light. The gaze will shift to the pedestrian crossing the street. This fixation guides the gaze shift to monitor the pedestrian's movement to ensure a safe response to the traffic signal change.

The scene shows a narrow city street with parked cars on both sides and a pedestrian walking on the sidewalk. The gaze is currently focused on the lead vehicle. The gaze will shift to the pedestrian. This fixation guides the gaze shift to assess the pedestrian's movement and ensure a safe response while maintaining awareness of the vehicle ahead.

The street scene shows a multi lane urban road with parked cars on both sides, a traffic light at the intersection, and several pedestrians on the sidewalks. The gaze is currently focused on the lead vehicle in the center lane. The gaze will shift to the pedestrian crossing the street ahead. This fixation guides the gaze shift to monitor the pedestrian's movement to ensure a safe yield at the intersection.

Fig. 9: FSDAM predictions on BDD-A test set. Ground truth gaze (left) vs. FSDAM predictions (right) with generated captions. Our model accurately identifies safety-critical regions and generates coherent explanations across diverse urban scenarios: intersections with pedestrians, narrow streets with parked vehicles, and multi-lane roads with complex traffic.



Number of regions: 2. Region 1: Road ahead. Region 2: Crossing pedestrians. Reason 1: To monitor traffic and obstacles directly in the path for safe navigation. Reason 2: To ensure the safety of pedestrians and be prepared to stop.

Number of regions: 2. Region 1: Cars parked along both sides of the street. Region 2: Pedestrians crossing the street at the crosswalk. Reason 1: To maintain a safe distance from parked cars. Reason 2: To yield to pedestrians crossing the street.

Number of regions: 1. Region 1: Vehicle ahead. Reason 1: To monitor the distance and behavior of the vehicle ahead to ensure safe following and react to any changes in its movement.

Number of regions: 1. Region 1: The blue car in the center lane. Reason 1: To monitor the blue car's position and speed to maintain a safe following distance and to be prepared for sudden braking or lane changes.

1. Number of regions: 3 2. Region 1: The vehicle directly ahead Region 2: The lane to the right Region 3: The road further ahead 3. Reason 1: To monitor the speed and distance from the vehicle directly ahead for safe deceleration. Reason 2: To check for any obstacles or vehicles in the lane to the right before completing the lane change. Reason 3: To anticipate upcoming traffic conditions and plan the driving path accordingly.

Number of regions: 3. Region 1: Left lane vehicle. Region 2: Lead vehicle in the center lane. Region 3: Intersection ahead. Reason 1: To monitor the left lane vehicle's position and speed for safe following. Reason 2: To track the lead vehicle in the center lane and maintain a stable driving path. Reason 3: To observe the intersection ahead and prepare for any required action.

Fig. 10: Zero-shot FSDAM predictions on W3D [62]. Ground truth annotations (left) and FSDAM predictions (right) highlight the same attention targets and safety cues across diverse scenes. FSDAM generalizes to W3D without any dataset specific training and produces gaze and reasoning patterns consistent with the labeled regions.

description, current fixation, anticipated shift, and safety rationale. Notably, the model correctly identifies gaze targets from its own predictions and provides appropriate driving context. For instance, when detecting pedestrians, it generates “monitor the pedestrian’s movement to ensure a safe yield" rather than generic responses. This consistency across varied scenarios from narrow streets to complex intersections demonstrates that our temporal mining captures fundamental attention patterns that generalize beyond the training distribution.

B.2 Cross-data generalization on W3D dataset

We evaluate FSDAM’s zero-shot transfer to W3D, which provides ground truth captions alongside gaze annotations. Figure 10 shows representative examples comparing our predictions with W3D’s region-based annotations.

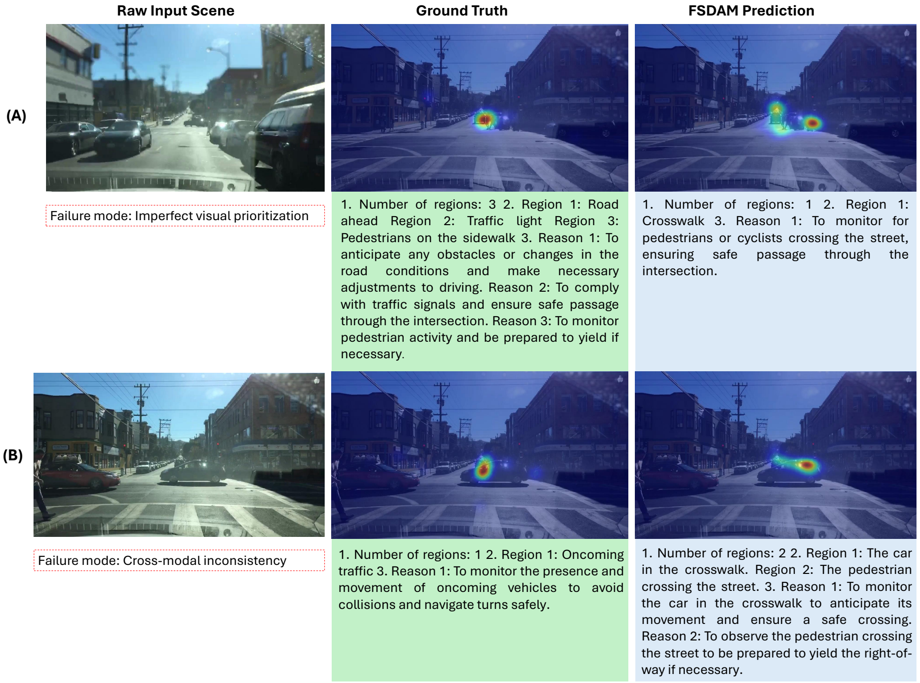
W3D uses a distinctive region-numbering format ("Number of regions: 2. Region 1: Road ahead...") that differs fundamentally from FSDAM’s output structures. Despite this, both output structures consistently identify identical attention targets across all examples. In pedestrian scenarios, both output emphasize crosswalk regions; in vehicle-following situations, they consistently track the lead vehicle; and at intersections, they distribute attention across several possible conflict zones.

The gaze heatmaps show strong spatial alignment, with FSDAM accurately predicting attention concentration at the same locations as ground truth.

The semantic correspondence extends to safety reasoning despite surface-level differences, where W3D states “ensure the safety of pedestrians and be prepared to stop", FSDAM generates “yield to pedestrians crossing the street"—expressing the same defensive driving principle. This alignment emerges without any W3D-specific training, suggesting that our 90-sample curation captures fundamental attention patterns that generalize across annotation protocols and datasets. Strong zero-shot performance validates our temporal mining approach for few-shot learning in safety-critical applications.

B.3 Failure Cases

To better characterize the limitations of FSDAM, we qualitatively analyzed failure cases along two axes: *imperfect visual prioritization* and *cross-modal inconsistency*. The first refers to cases in which the predicted attention map remains partially grounded in the scene but misprioritizes the most behaviorally relevant cue or becomes overly diffuse under visual clutter. The second refers to cases in which the generated explanation is plausible at the language level but is not tightly supported by the attended visual evidence. This taxonomy is useful because it distinguishes failures of spatial allocation from cross-modal inconsistency between the gaze prediction and the reasoning, which need not occur simultaneously. In several cases, the model identifies the correct high-level maneuver context while still allocating attention to suboptimal regions. In others, the attention map and explanation each appear locally plausible, yet they do not support the same driving rationale when considered jointly. These examples


Raw Input Scene
Ground Truth
FSDAM Prediction
(A)
Failure mode: Imperfect visual prioritization
(B)
Failure mode: Cross-modal inconsistency

1. Number of regions: 3 2. Region 1: Road ahead Region 2: Traffic light Region 3: Pedestrians on the sidewalk 3. Reason 1: To anticipate any obstacles or changes in the road conditions and make necessary adjustments to driving. Reason 2: To comply with traffic signals and ensure safe passage through the intersection. Reason 3: To monitor pedestrian activity and be prepared to yield if necessary.

1. Number of regions: 1 2. Region 1: Crosswalk 3. Reason 1: To monitor for pedestrians or cyclists crossing the street, ensuring safe passage through the intersection.

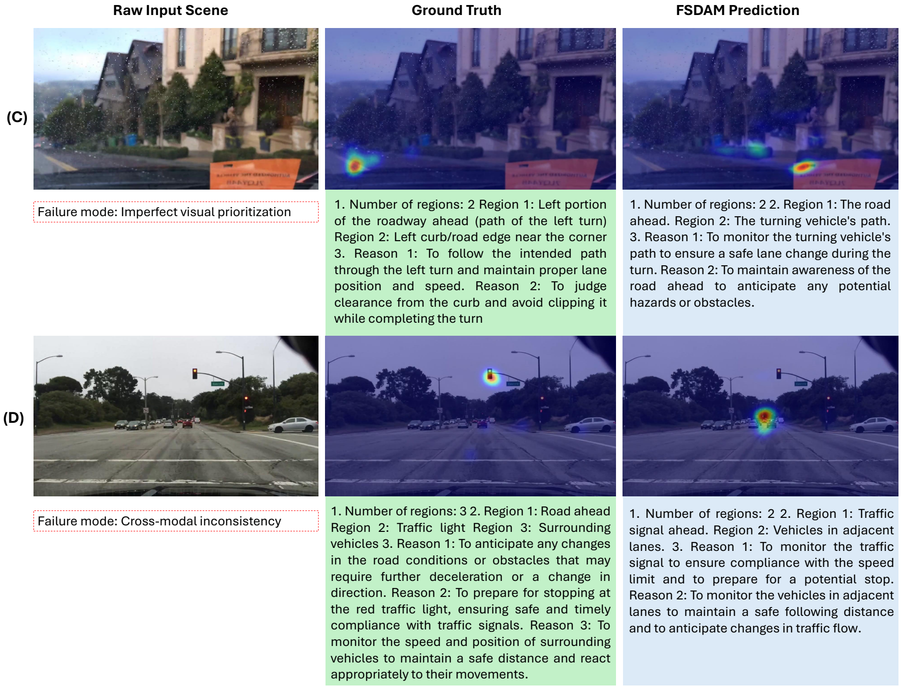
1. Number of regions: 1 2. Region 1: Oncoming traffic 3. Reason 1: To monitor the presence and movement of oncoming vehicles to avoid collisions and navigate turns safely.

1. Number of regions: 2 2. Region 1: The car in the crosswalk. Region 2: The pedestrian crossing the street. 3. Reason 1: To monitor the car in the crosswalk to anticipate its movement and ensure a safe crossing. Reason 2: To observe the pedestrian crossing the street to be prepared to yield the right-of-way if necessary.

Fig. 11: Qualitative failure cases of FSDAM. **(A) Imperfect visual prioritization:** the model partially captures the crossing context but also assigns substantial attention to a visually salient right-side object, diluting focus on the primary crossing-related cue. **(B) Cross-modal inconsistency:** the explanation emphasizes crossing-related hazards, whereas the predicted attention is not aligned with the reference oncoming-conflict region.

therefore help clarify not only where FSDAM fails, but also how those failures arise.

Figure 11 shows two representative failures. In Fig. 11(A), the model attends to part of the relevant crossing region but also exhibits spurious allocation to a secondary salient object, indicating imperfect prioritization within the scene.This suggests that the model is sensitive to conspicuity even when a more behaviorally important target is present nearby. The error is therefore not a complete loss of scene understanding, but a failure to rank competing cues correctly. Such cases are especially important in urban scenes, where multiple salient objects co-occur but only a subset directly constrains the driving decision. In Fig. 11(B), the error is cross-model inconsistency: although the generated explanation refers to plausible crossing-related entities, the predicted attention does not support the same hazard configuration as the reference map. This suggests that explanation quality can remain superficially plausible even when visual grounding is incomplete. More broadly, the example indicates that textual plausibility alone is not sufficient evidence of faithful reasoning.



Failure mode: Imperfect visual prioritization

1. Number of regions: 2 Region 1: Left portion of the roadway ahead (path of the left turn) Region 2: Left curb/road edge near the corner 3. Reason 1: To follow the intended path through the left turn and maintain proper lane position and speed. Reason 2: To judge clearance from the curb and avoid clipping it while completing the turn

1. Number of regions: 2 2. Region 1: The road ahead. Region 2: The turning vehicle's path. 3. Reason 1: To monitor the turning vehicle's path to ensure a safe lane change during the turn. Reason 2: To maintain awareness of the road ahead to anticipate any potential hazards or obstacles.

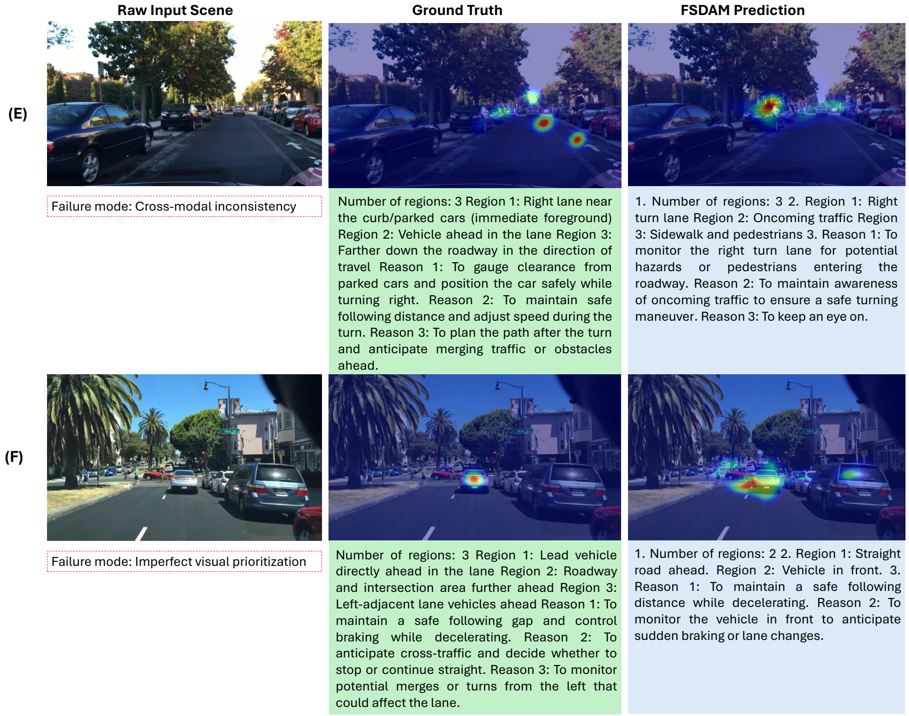
Failure mode: Cross-modal inconsistency

1. Number of regions: 3 2. Region 1: Road ahead Region 2: Traffic light Region 3: Surrounding vehicles 3. Reason 1: To anticipate any changes in the road conditions or obstacles that may require further deceleration or a change in direction. Reason 2: To prepare for stopping at the red traffic light, ensuring safe and timely compliance with traffic signals. Reason 3: To monitor the speed and position of surrounding vehicles to maintain a safe distance and react appropriately to their movements.

1. Number of regions: 2 2. Region 1: Traffic signal ahead. Region 2: Vehicles in adjacent lanes. 3. Reason 1: To monitor the traffic signal to ensure compliance with the speed limit and to prepare for a potential stop. Reason 2: To monitor the vehicles in adjacent lanes to maintain a safe following distance and to anticipate changes in traffic flow.

Fig. 12: Qualitative failure cases of FSDAM. **(C) Imperfect visual prioritization:** during a turning maneuver, attention is diverted toward a foreground dashboard artifact rather than the turn path and curb-clearance region emphasized by the reference map. **(D) Cross-modal inconsistency:** the explanation mentions the traffic signal and surrounding vehicles, but the predicted attention is concentrated more strongly on the lead vehicle and forward roadway than on the signal highlighted in the reference.

Figure 12 demonstrates that these failure modes recur in visually distinct settings. In Fig. 12(C), the dominant error is spatial misprioritization caused by foreground saliency. Here, the model is distracted by an image-plane artifact that is visually prominent but irrelevant to the maneuver, revealing limited robustness to nuisance saliency. This type of error is noteworthy because it arises even when the semantically relevant cue is spatially localized and structurally simple. It suggests that improved suppression of foreground bias or stronger supervision on maneuver-critical regions may be beneficial. In Fig. 12(D), the model again produces an explanation that is semantically plausible but only partially supported by the attended region. The explanation refers to the traffic signal and nearby traffic, whereas the heatmap is dominated by the lead vehicle and forward roadway. Together, these cases indicate that FSDAM remains sensitive both to irrelevant saliency and to incomplete grounding between attention prediction and textual reasoning.


Raw Input Scene
Ground Truth
FSDAM Prediction
(E)
Failure mode: Cross-modal inconsistency
(F)
Failure mode: Imperfect visual prioritization

Number of regions: 3 Region 1: Right lane near the curb/parked cars (immediate foreground) Region 2: Vehicle ahead in the lane Region 3: Farther down the roadway in the direction of travel Reason 1: To gauge clearance from parked cars and position the car safely while turning right. Reason 2: To maintain safe following distance and adjust speed during the turn. Reason 3: To plan the path after the turn and anticipate merging traffic or obstacles ahead.

1. Number of regions: 3 2. Region 1: Right turn lane Region 2: Oncoming traffic Region 3: Sidewalk and pedestrians 3. Reason 1: To monitor the right turn lane for potential hazards or pedestrians entering the roadway. Reason 2: To maintain awareness of oncoming traffic to ensure a safe turning maneuver. Reason 3: To keep an eye on.

Number of regions: 3 Region 1: Lead vehicle directly ahead in the lane Region 2: Roadway and intersection area further ahead Region 3: Left-adjacent lane vehicles ahead Reason 1: To maintain a safe following gap and control braking while decelerating. Reason 2: To anticipate cross-traffic and decide whether to stop or continue straight. Reason 3: To monitor potential merges or turns from the left that could affect the lane.

1. Number of regions: 2 2. Region 1: Straight road ahead. Region 2: Vehicle in front. 3. Reason 1: To maintain a safe following distance while decelerating. Reason 2: To monitor the vehicle in front to anticipate sudden braking or lane changes.

Fig. 13: Qualitative failure cases of FSDAM. **(E) Cross-modal inconsistency:** in this right-turn scene, the explanation refers to maneuver-relevant hazards, but the attended regions do not align with the reference cues associated with curb clearance, lead-vehicle monitoring, and downstream path planning. **(F) Imperfect visual prioritization:** in a cluttered multi-agent scene, the model captures the general forward driving context but distributes attention too broadly across nearby agents and scene elements relative to the more concentrated reference map.

Figure 13 highlights two further limitations. In Fig. 13(E), the mismatch is again cross-modal: the explanation describes relevant turning hazards, but the visual evidence emphasized by the model does not correspond to the same set of cues. This failure is informative because the explanation is not nonsensical; rather, it is insufficiently grounded in the actual attended regions. In Fig. 13(F), the prediction remains broadly grounded in the correct forward scene but loses precision under clutter, yielding a diffuse allocation over multiple agents and scene elements. The error here is therefore one of concentration rather than semantic collapse. This distinction matters, since it suggests that part of the remaining performance gap may stem from uncertainty calibration or attention sharpening rather than from missing scene semantics altogether. It also indicates that cluttered multi-agent scenes remain a particularly challenging regime for the model.

Table 5: Comparison of captioning performance across driving scenarios on W3D dataset. Fully-trained baselines use the original W3D training data, Zero-shot and ICL models require no fine-tuning, and our FSDAM and few-shot baselines are trained on a 90-sample BDD-A subset. Higher is better.

Method	**Training Regime**	**Normal Driving**				**Safety-Critical Situation**				**Traffic Accident**			
		BLEU	METEOR	ROUGE	CIDEr-R	BLEU	METEOR	ROUGE	CIDEr-R	BLEU	METEOR	ROUGE	CIDEr-R
Fully Trained on W3D													
GazeXplain* [8]	Full-data (W3D) [∼70k samples]	0.31	0.30	0.22	0.42	0.19	0.29	0.37	0.55	0.17	0.20	0.44	0.66
LLada* [62]	Full-data (W3D) [∼70k samples]	**0.44**	0.36	**0.58**	**0.96**	**0.44**	**0.38**	**0.59**	**1.23**	**0.38**	0.32	**0.52**	**1.00**
Zero-shot and In-Context Models													
Qwen-VL [3]	Zero-shot (no training)	0.10	0.19	0.28	0.34	0.19	0.21	0.29	0.13	0.08	0.21	0.29	0.12
LLaVA [36]	Zero-shot (no training)	0.12	0.14	0.23	0.35	0.13	0.19	0.11	0.10	0.17	0.26	0.19	0.13
Qwen-VL [3]	In-context learning (no fine-tuning)	0.13	0.18	0.22	0.36	0.21	0.17	0.30	0.23	0.12	0.24	0.33	0.15
Few-shot Learning (BDD-A 90 samples)													
DeepGazeI [27] + LLaVA	Few-shot	0.12	0.23	0.28	0.14	0.13	0.22	0.30	0.18	0.15	0.21	0.31	0.17
DeepGazeIIE [34] + LLaVA	Few-shot	0.11	0.18	0.26	0.17	0.11	0.20	0.32	0.13	0.11	0.19	0.34	0.10
MLNet [10] + LLaVA	Few-shot	0.13	0.19	0.27	0.31	0.26	0.20	0.32	0.12	0.13	0.18	0.33	0.28
LLada† [62]	Few-shot	0.15	0.18	0.19	0.12	0.11	0.16	0.16	0.10	0.13	0.18	0.19	0.17
FSDAM (Ours)	Few-shot	0.42	**0.37**	0.48	0.83	0.35	0.33	0.46	0.47	0.33	**0.35**	0.34	0.84

Overall, these examples show that FSDAM failures arise primarily from two sources: inaccurate prioritization among competing visual cues and incomplete alignment between attention maps and generated explanations. Across cases, the model often captures the global maneuver context correctly, but does not always localize the most behaviorally critical cue with sufficient precision. The qualitative evidence also suggests that explanation generation can abstract the scene at the correct semantic level while still failing to remain faithful to the underlying attended evidence. This gap between semantic plausibility and visual faithfulness is especially important for explainable driving models, where interpretability is intended to reflect the model's actual decision basis. From a modeling perspective, the observed errors point toward three concrete directions: improved suppression of nuisance saliency, stronger supervision for maneuver-critical cue selection, and tighter coupling between explanation generation and attention prediction. Taken together, the failure analysis suggests that the main challenge is not recognizing the scene at a coarse level, but resolving which cue should dominate attention and how that cue should be verbalized consistently.

C Additional Quantitative Analysis

C.1 Extension of Table 3: Adding LLada in Few-Shot Setting

To provide a more rigorous comparison, we extend Table 3 of the main paper by including LLada [62] fine-tuned under the same few-shot regime as our method. Specifically, LLada† is fine-tuned on the identical 90-sample BDD-A subset used to train FSDAM, ensuring a controlled and fair comparison. As shown in Table 5, LLada† under the few-shot setting performs substantially below its fully-trained counterpart, highlighting the difficulty of adapting language models for driving to extremely limited supervision. In contrast, FSDAM consistently achieves competitive performance across all three scenario categories despite using the same limited training data, demonstrating the effectiveness of our dual-pathway architecture and blur-gap regularization in low-data regimes.

Table 6: Variance analysis of FSDAM on BDD-A across three random seeds (90 training samples). Results are reported as mean ± standard deviation.

Method	CC↑	KL↓	SIM↑	AUC-J↑	AUC-B↑	NSS↑
FSDAM (Seed 1)	0.60	1.13	0.43	0.96	0.91	4.10
FSDAM (Seed 2)	0.58	1.17	0.42	0.95	0.90	3.98
FSDAM (Seed 3)	0.61	1.11	0.44	0.96	0.91	4.15
Mean ± Std	**0.60 ± 0.01**	**1.14 ± 0.03**	**0.43 ± 0.01**	**0.96 ± 0.01**	**0.91 ± 0.01**	**4.08 ± 0.09**

C.2 Variance Analysis Under Random Support Sampling

To evaluate the stability of FSDAM under different support-set compositions, we report performance on BDD-A across three random seeds and report mean ± standard deviation for each metric. Table 6 shows that performance remains consistent across seeds, with only small fluctuations across all metrics. This indicates that FSDAM is not sensitive to the specific random selection of few-shot support samples, and that the observed improvements do not depend on a particularly favorable support-set draw. These results confirm that FSDAM's performance is reproducible and robust to support-set composition in the low-data regime.