S.I. Object Detection and Uncertainty Calculation

To obtain 2D object and uncertainty information, we first train an object detection model YOLO V8, on the PeSOTIF dataset using an 8:2 split for training and valid sets. The results are summarized in Table I. As can be seen, YOLOV8m achieves high precision and recall when compared to other models.

TABLE I: Performance Comparison on PeSOTIF Valid Set

Model	mAP@0.5	Recall	mAP@0.5:0.95
YOLOV8n	0.503	0.477	0.285
YOLOV8s	0.534	0.49	0.304
YOLOV8m	**0.57**	**0.548**	**0.321**
YOLOV8l	0.543	0.522	0.311
YOLOV9c	0.535	0.503	0.303

Building on the work proposed in [1], we ensemble five YOLO8m [2] models and apply the Basic Sequential Algorithm Scheme with intra-sample exclusivity (BSAS excl.) using Intersection over Union (IOU) to create detection result clusters. Each cluster contains unique samples from different models. Next, we remove redundant detection results across different clusters using the Non-Maximum Suppression (NMS) method based on type, 2D bounding box position, size, and confidence scores.

To quantify the uncertainty level, we integrate the SOTIF entropy method proposed in [3]. Specifically, the probability of class c, p_c is calculated as:

$$p_c = p(y = c \mid x, D) \approx \frac{1}{T} \sum_{t=1}^{T} p(y = c \mid x, W_t) \tag{1}$$

which defines p_c as the probability of an input x being classified as class c given the dataset D. This is approximated by averaging the probabilities over T different model instances, each with a different weight W_t.

SOTIF entropy is calculated as:

$$E_{pe} = -\sum_{c} \left(p_c \log p_c + (1 - p_c) \log(1 - p_c)\right) \tag{2}$$

where $p_c \log p_c$ represents entropy contribution from predicted class probabilities, while the other term accounts for uncertainty in other classes' probabilities.

To penalize missed or ghost detections, the final perception SOTIF entropy, E^*_{pe}, is scaled by the number of networks that did not detect the object, as shown below:

$$E^*_{pe} = E_{pe} \left(1 + f_p(T - d)\right) \tag{3}$$

where: f_p is the penalty factor; T is the total number of networks; d is the number of networks that detected the object. The uncertainty level is converted to semantic texts using three entropy levels defined in [3]:

$$l_u = \begin{cases} 0 & \text{if } E^*_{pe} < \theta_{lm} \quad \text{(low, normal)} \\ 1 & \text{if } \theta_{lm} \le E^*_{pe} < \theta_{mh} \quad \text{(medium, caution)} \\ 2 & \text{if } \theta_{mh} \le E^*_{pe} \quad \text{(high, warning)} \end{cases} \tag{4}$$

where l_u represents the uncertainty level, θ_{lm} refers to the uncertainty threshold between low and medium uncertainty, and θ_{mh} denotes the uncertainty threshold between medium and high uncertainty. Parameters used in this section are shown in Table II.

TABLE II: Parameters for Uncertain Information

Parameter	Description	Value
T	Number of model instances	5
W_t	Weights of t-th model instance	1.0
f_p	Penalty factor	0.1
θ_{lm}	Uncertainty threshold between low and medium uncertainty	1.2
θ_{mh}	Uncertainty threshold between medium and high uncertainty	1.6

S.II. Fine-tuning Parameters

TABLE III: Settings for Fine-tuning and Evaluation

Parameter	Value	Parameter	Value
Image Captioning			
Max Epoch	5	LR Scheduler	Linear Warmup Cos
Max Length	50	Initial LR	1e-5
Min Length	8	Minimum LR	0
Number of Beams	5	Warmup LR	1e-8
Image Size	364	Warmup Steps	1000
Freeze ViT	True	Weight Decay	0.05
VQA			
Learning Rate	5e-5	Max New Tokens	512
Cutoff Length	1024	Temperature	0.2/0.7
Top-p	0.7	Quantization	BitsandBytes

S.III. Dataset Generation and Validaiton

We show the prompts used for the dataset generation, including generation/validation prompts as well as the final verification, in Figure 1. The LLM-as-a-judge prompt used for the open-ended VQA results assessment is presented in Figure 2. The average score is calculated as the mean of the four criteria scores, and the final score is the average of all model scores.

S.IV. Human Evaluation

A research team member manually validated a sample of 595 entries from the dataset to assess quality and accuracy. We determined the sample size to use with the finite population formula:

$$n = \frac{NZ^2p(1-p)}{E^2(N-1) + Z^2p(1-p)}$$

where $N = 5580$ (dataset size), $Z = 2.58$ (99% confidence level), $p = 0.5$ (maximum variance), and $E = 0.05$ (margin of error). Substituting these values, we obtain $n \approx 594.79$, which we round up to 595.

Prompt:

We are studying perception problems of Safety of the Intended Functionality (SOTIF) problems. Generate a high-quality caption for the image uploaded. Be objective, natural, and human-like. The caption should focus on SOTIF-relevant elements, including:

- Traffic scenario description
- Environmental conditions (weather, lighting, visibility)
- Road users and their behaviors
- Infrastructure elements
- Any unusual or edge-case elements

Provide:

- A comprehensive caption (2–4 sentences)
- Confidence score (1–5, where 5 is most confident)
- Brief explanation of SOTIF relevance
- Key elements list (important objects/conditions identified)

(a) Caption generation prompt

Prompt:

We are building a dataset for Safety of the Intended Functionality (SOTIF) research in autonomous vehicles. Analyze this caption for the given traffic scene image:
Caption: ”caption”
Evaluate:

1) Factual accuracy: Does the caption correctly describe what you see?
2) Completeness: Are important safety-relevant elements mentioned?
3) SOTIF relevance: Does it capture elements relevant for AV safety analysis?
4) Clarity and precision of description

Rate accuracy (1-5 scale) and identify any issues. If issues exist, suggest specific improvements.

(b) Caption validation prompt

Prompt:

Generate safety-critical questions for this traffic scene image based on SOTIF analysis. Focus on potential edge cases and safety concerns.
Image Caption: ”caption”
Generate questions about:

- Environmental hazards and visibility issues
- Road user behavior and interactions
- Infrastructure limitations or anomalies
- Potential system failures or edge cases

Provide:

- 3-5 relevant safety questions
- Question difficulty level (1-5)
- Expected answer type for each question

(c) Question generation prompt

Prompt:

Validate the generated questions for this traffic scene:
Questions: ”questions”
Evaluate:

1) Relevance: Are questions pertinent to SOTIF analysis?
2) Safety focus: Do they address genuine safety concerns?
3) Clarity: Are questions well-formulated and unambiguous?
4) Difficulty appropriateness: Is complexity suitable for the scene?

Rate overall quality (1-5 scale) and suggest improvements.

(d) Question validation prompt

Prompt:

Answer the following safety question about this traffic scene image. Base your response on SOTIF principles and autonomous vehicle safety considerations.
Image Caption: ”caption” Question: ”question”
Provide:

- Detailed answer addressing safety implications
- Risk assessment (low/medium/high)
- Potential mitigation strategies
- Confidence level in response (1-5)

(e) Answer generation prompt

Prompt:

Validate this answer to the safety question:
Question: ”question” Answer: ”answer”
Evaluate:

1) Accuracy: Is the answer factually correct and realistic?
2) Completeness: Does it address all aspects of the question?
3) Safety relevance: Are SOTIF principles properly applied?
4) Practicality: Are suggestions feasible and actionable?

Rate answer quality (1-5 scale) and identify improvements.

(f) Answer validation prompt

Prompt:

Perform overall validation of the complete dataset entry:
Caption: ”caption” Questions: ”questions” Answers: ”answers”
Evaluate the entire entry for:

1) Coherence: Do all components work together logically?
2) SOTIF alignment: Does the entry support safety analysis objectives?
3) Quality consistency: Are all components of similar high quality?
4) Dataset value: Will this entry contribute meaningfully to research?

Provide overall rating (1-5) and final recommendations for dataset inclusion.

(g) Overall validation prompt

Fig. 1: Prompts used for dataset generation and validation workflow

Prompt:

Instructions:
You are an impartial judge evaluating the performance of an AI model's response to a SOTIF (Safety of the Intended Functionality) question for autonomous driving.
user prompt = Question: (question)
Reference Answer: (reference-answer)
Model Answer: (model-answer)
Image Context: (image-description)
Please evaluate the Model Answer against the Reference Answer using the 5-point rubric. For each criterion, provide both a numerical score (1-5) and detailed reasoning for that score.

1) **Relevance**
How closely and directly does the model answer the specific SOTIF question?
   - 1: Largely off-topic or irrelevant
   - 2: Minimally relevant; significant deviations
   - 3: Moderately relevant; some unrelated content
   - 4: Highly relevant; minor deviations
   - 5: Perfectly relevant; fully aligned

2) **Trustworthiness**
How accurate, reliable, and safe is the information?
   - 1: Highly unreliable; factual errors/harmful content
   - 2: Minimally trustworthy; several inaccuracies
   - 3: Generally accurate; some minor errors
   - 4: Highly trustworthy; negligible errors
   - 5: Flawless accuracy; fully compliant

3) **Clarity**
How easy is it to understand the explanation?
   - 1: Incomprehensible
   - 2: Poor clarity; frequent ambiguities
   - 3: Adequate; some ambiguities
   - 4: Clear; minor issues
   - 5: Crystal-clear; exemplary articulation

4) **Coherence**
How logical and well-structured is the response?
   - 1: Disjointed; lacks logical flow
   - 2: Poor coherence; logical gaps
   - 3: Some logical flow; occasional inconsistencies
   - 4: Highly coherent; minor inconsistencies
   - 5: Perfectly cohesive; seamless flow

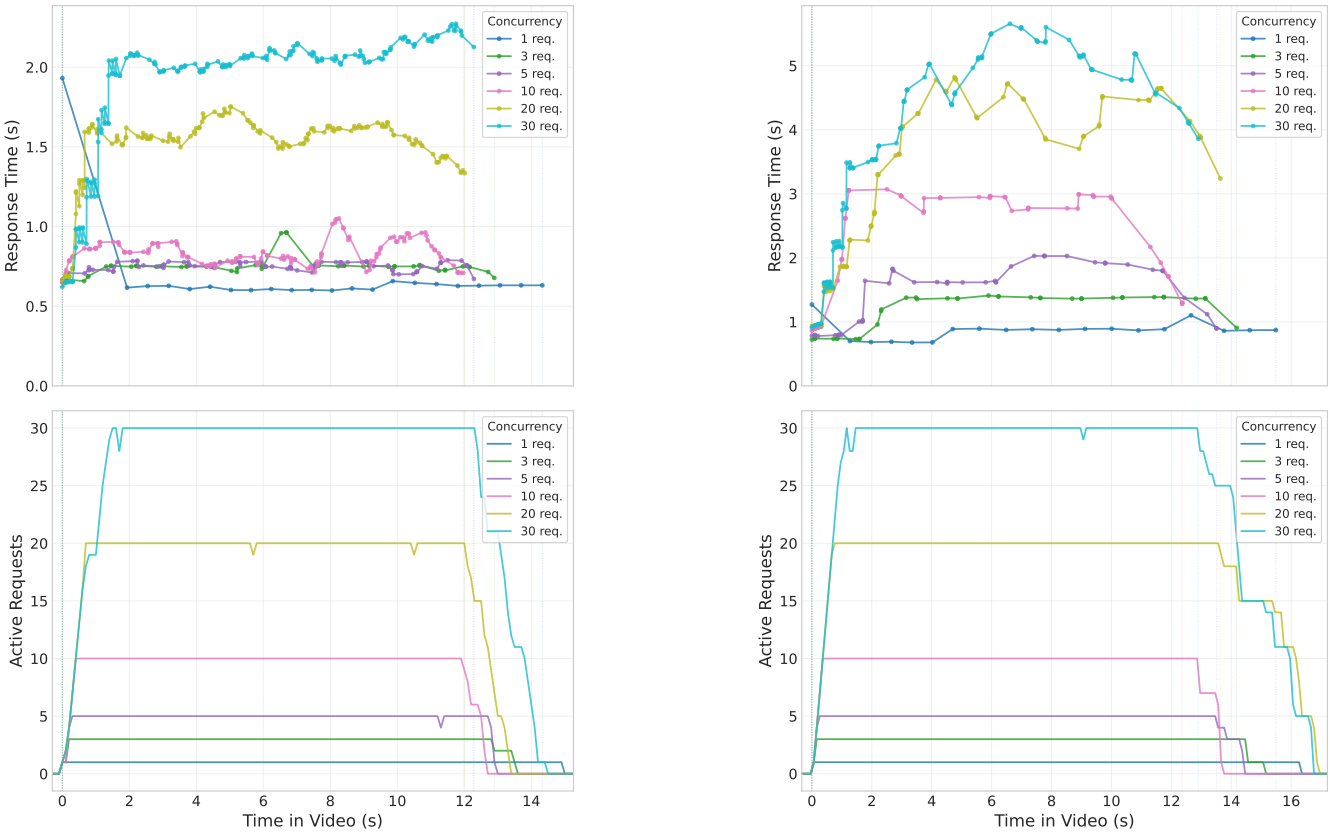
Fig. 2: LLM-as-a-judge prompt for open-ended VQA results assessment

S.V. Additional dynamic processing results

Results from the RTX 4090 GPU are shown in Figures 3a and 3c, while results from the dual RTX 4090 GPU setup are presented in Figures 3b and 3d. Interestingly, the dual RTX 4090 configuration exhibits higher response time and similar queue waiting time compared to the single RTX 4090. This is likely due to PCIe bandwidth limitations: when both GPUs are installed, each is limited to PCIe 4.0 x8, rather than the full x16. In contrast, datacenter GPUs are often connected via NVLink, which offers significantly higher bandwidth and lower latency for inter-GPU communication compared to PCIe.

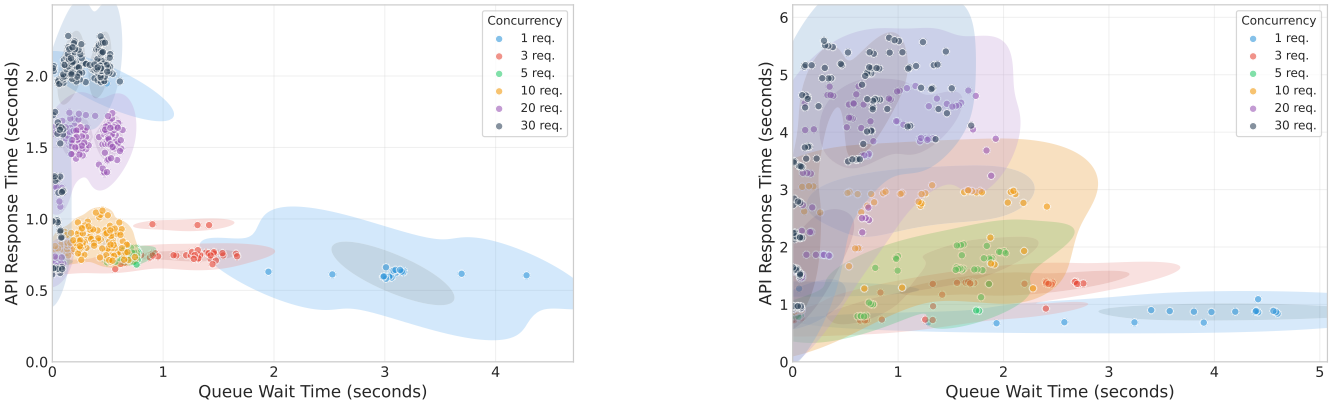
References

[1] L. Peng, J. Li, W. Shao, and H. Wang, "Pesotif: A challenging visual dataset for perception sotif problems in long-tail traffic scenarios," in *2023 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2023, pp. 1–8.
[2] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLO," Jan. 2023. [Online]. Available: https://github.com/ultralytics/ultralytics
[3] L. Peng, B. Li, W. Yu, K. Yang, W. Shao, and H. Wang, "Sotif entropy: Online sotif risk quantification and mitigation for autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, 2023.



(a) Single RTX 4090

(b) Dual RTX 4090



(c) Single RTX 4090

(d) Dual RTX 4090

Fig. 3: Comparison of continuous inference performance on single vs. dual RTX 4090. Top row: Average generation time per image. Bottom row: Tradeoff between API response time and queue waiting time.