Model-Agnostic Uncertainty-Aware Semantic Segmentation with Conformal Risk Guarantees for Scene Understanding

Bakary Badjie
LASIGE, Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa, Portugal

José Cecílio
LASIGE, Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa, Portugal

Nils-Jonathan Friedrich
Technische Universität Bergakademie Freiberg, Germany

Norman Seyffer
Technische Universität Bergakademie Freiberg, Germany

Georg Jäger
Technische Universität Bergakademie Freiberg, Germany

António Casimiro
LASIGE, Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa, Portugal

Abstract

Accurate and reliable scene segmentation is a fundamental requirement for autonomous navigation systems operating in open and dynamic environments. As these systems increasingly rely on data-driven perception modules, their safety and operational robustness hinge on well-calibrated uncertainty estimates that can support explicit control of prediction errors through conformal calibration. Most existing uncertainty-aware segmentation approaches remain architecture-specific and are not evaluated under a common uncertainty-and-calibration protocol across distinct segmentation architectures and datasets. This work introduces a model-agnostic conformal segmentation pipeline that enables operationally meaningful, calibration-based error control in real-world deployments. The proposed framework treats segmentation networks as black boxes and operates on per-pixel class probabilities that are fine-tuned through evidential deep learning (EDL) to decompose aleatoric and epistemic uncertainties. We then apply pixel-wise, class-conditional split-conformal calibration to derive acceptance thresholds for user-defined target error rates. We instantiate the pipeline with DINOv2, Mask2Former, and SegFormer and evaluate it on a newly collected Lisbon street scene (LiSS) dataset; additional cross-dataset results on COCO, using a restricted set of safety-relevant classes, are reported in the appendix. Results show architecture- and class-dependent in-domain uncertainty-error alignment and indicate that dataset shift weakens uncertainty-based filtering and conformal risk control. This motivates continuous monitoring and recalibration as a practical requirement for trustworthy segmentation in safety-critical navigation.

2012 ACM Subject Classification Software and its engineering → Software reliability; Computing methodologies → Computer vision; Computer systems organization → Reliability

Keywords and phrases semantic segmentation, uncertainty quantification, evidential deep learning, conformal prediction, risk control, selective prediction

Digital Object Identifier 10.4230/OASIcs.AEiC.2026.1

Funding *Bakary Badjie*: Supported by FCT Ph.D. grant agreement No. 2023.02832.BD (LASIGE UID/00408/2025, DOI: `https://doi.org/10.54499/UID/00408/2025`).
José Cecílio: This work was supported by the Fundação para a Ciência e a Tecnologia (FCT) under the RobustSAR project (ref. 2024.07838.CBM) and the LASIGE Research Unit (ref. UID/00408/2025, DOI: `https://doi.org/10.54499/UID/00408/2025`).
Georg Jäger: This work was partially funded by the German Federal Ministry for Education and Research within the DAAD program (Project RobustSAR with ID 57764655).

1 Introduction

Reliable scene understanding is essential for autonomous systems operating in dynamic, safety-critical environments. Semantic segmentation is the core component of visual perception systems for autonomous driving, robotics, and urban monitoring. It provides dense scene understanding at the pixel level [5]. Recent works have shown that transformer-based architectures achieve strong performance on large-scale benchmark datasets. Examples include Mask2Former for universal image segmentation [4], SegFormer for efficient transformer-based semantic segmentation [19], and self-supervised vision backbones such as DINOv2 and DINOv3 that transfer well to dense prediction tasks [16, 18]. These advances have improved accuracy on standard datasets, but they do not in themselves quantify when predictions can be trusted in safety-critical settings.

In operational settings, perception errors can lead to unsafe decisions, and performance often deteriorates under domain shifts or adverse conditions. To support trustworthy autonomy, segmentation models must provide uncertainty estimates that meaningfully reflect risk and allow explicit error control. Uncertainty quantification for dense prediction has therefore received increasing research attention [10]. Approaches such as Monte Carlo dropout (MC-Dropout), deep ensembles, and evidential deep learning (EDL) have been extended to semantic segmentation in both natural and medical domains [1, 8, 9]. Evidential methods are particularly attractive because they provide estimates of both aleatoric and epistemic uncertainty in a single forward pass [7]. However, most existing studies focus on specific network architectures or application domains [9, 17, 20]. As a result, there is limited evidence on how uncertainty estimation techniques behave across heterogeneous segmentation models and datasets.

From a deployment perspective, accurate uncertainty scores are not sufficient [13]. Practitioners need guarantees that relate uncertainty to error probabilities. Conformal prediction provides a distribution-free way to construct prediction sets with finite-sample coverage guarantees. It has recently been adapted to segmentation and other dense prediction problems [15, 3, 12]. These works demonstrate that conformal prediction can highlight unreliable regions in real-world segmentation and complement calibration-based approaches. Yet they typically target a single model type or a specific scenario. There remains a gap in understanding how conformal prediction interacts with different forms of uncertainty estimates and whether a single pipeline can be applied across architectures and datasets without model-specific modifications.

In this work, we propose an architecture-independent uncertainty-aware segmentation pipeline with class-conditional conformal risk control. The pipeline operates on per-pixel class probability maps produced by a segmentation model and does not depend on the internal backbone or decoder design. The methodological contribution is therefore not a new segmentation architecture but a unified post-training pipeline that can be applied without architecture-specific redesign and evaluated under a common risk-control protocol across models and datasets. We use evidential deep learning (EDL) to fine-tune the model and obtain aleatoric and epistemic uncertainty estimates from Dirichlet predictions [1, 8].

The contributions of this paper are as follows:

1. **Architecture-independent uncertainty and risk-control pipeline.** We introduce a unified post-training pipeline for semantic segmentation that operates on per-pixel class probability maps and is independent of the internal segmentation architecture. This yields a common interface for uncertainty estimation and selective prediction across different model architectures.

2. **Evidential fine-tuning for decomposed uncertainty.** We integrate EDL to obtain Dirichlet predictive distributions over per-pixel class probabilities and derive aleatoric and epistemic uncertainty scores for selective filtering.
3. **Pixel-wise, class-conditional conformal risk control.** We formulate a class-conditional conformal calibration procedure for pixel-wise uncertainty scores, yielding acceptance thresholds that target user-specified error levels and provide operational risk control on accepted predictions. Predictions that exceed the calibrated threshold are rejected rather than forced into a semantic class.
4. **Controlled cross-model and cross-dataset evaluation.** We instantiate the proposed evaluation pipeline on DINOv2, Mask2Former, and SegFormer on the LiSS and COCO datasets under a common protocol, enabling a direct analysis of how uncertainty quality and conformal risk control vary across architectures and datasets.

This work advances the reliability of perception systems by linking uncertainty estimation with explicit risk control. The proposed framework enables model-agnostic, interpretable confidence calibration, a requirement for safe operation across domains. By combining uncertainty estimation with conformal calibration, it provides empirically verifiable risk control under the assumptions of split-conformal calibration and thereby supports more trustworthy segmentation in real-world deployments.

2 Related Work

Recent progress on semantic segmentation, uncertainty quantification, and conformal prediction provides the context for our model-agnostic uncertainty and risk-control pipeline. Transformer-based segmentation architectures, including Mask2Former and SegFormer, achieve strong performance on benchmark datasets, and self-supervised backbones such as DINOv2 transfer effectively to dense prediction tasks [4, 19, 5, 16]. These works primarily optimize accuracy metrics such as mean Intersection over Union (mIoU) and typically do not provide explicit mechanisms for uncertainty quantification or risk control under domain shift (i.e., changes between the data distribution used for calibration or training and the data encountered at test time).

Uncertainty estimation for segmentation has been studied through Bayesian approximations, ensembles, and evidential approaches. Evidential deep learning (EDL) generates Dirichlet output distributions and supports decompositions of aleatoric and epistemic uncertainty in dense prediction [1, 8, 9, 7]. Recent works extend the EDL technique by combining evidential objectives with task- or modality-specific designs, including medical and LiDAR segmentation settings [17, 20]. However, existing evaluations of EDL often focus on a specific backbone or domain, and uncertainty measures are frequently used for qualitative inspection or ranking rather than for explicit, quantitative risk guarantees.

Conformal prediction is a post hoc calibration framework that converts model scores into prediction sets or acceptance rules with finite-sample validity under exchangeability. It has been adapted to segmentation pipelines to produce pixel-wise prediction sets and quality-control signals [15, 12, 3].

Distribution-free means that the conformal guarantee does not depend on a parametric assumption about the data-generating distribution and instead relies on exchangeability between calibration and test samples. Many pipelines assume a single architectural paradigm and a fixed nonconformity score. They often rely on scalar uncertainty proxies without separating epistemic and aleatoric components. As a result, the interaction between uncertainty estimation and conformal calibration across heterogeneous segmentation architectures remains insufficiently characterized.

Prior works have studied evidential uncertainty estimation and conformal prediction for segmentation, but typically within a single architectural family, application setting, or nonconformity design [4, 19, 16]. The main methodological difference from prior segmentation-specific uncertainty pipelines is that our method fixes the uncertainty representation and the conformal calibration protocol while varying the underlying segmentation backbone. Prior approaches usually couple the uncertainty mechanism to a specific architecture, training setting, or application domain. In contrast, our pipeline is defined at the level of per-pixel predictive probabilities and therefore separates risk-control analysis from model-specific design. Our pipeline supports a controlled comparison across different model architectures and datasets.

3 Methodology

This section describes the proposed model-agnostic pipeline for uncertainty-aware semantic segmentation and risk-controlled prediction. For uncertainty-aware semantic segmentation with risk-controlled selective prediction, we propose a two-part methodology: (i) a *design-time process* (cf. Figure 1) for uncertainty-aware prediction and conformal calibration, and (ii) a *run-time process* (cf. Figure 2) for risk-controlled selective prediction. The proposed method is architecture-independent because it uses only the model's per-pixel class probability maps as input to uncertainty estimation and conformal calibration.

Let C denote the number of semantic classes, H and W the image height and width and $u \in \Omega$ a pixel location in the image domain Ω. Let f_θ denote a segmentation model with parameters θ, and let $\alpha \in (0, 1)$ denote the target risk level used in conformal calibration.

At design time, we consider a segmentation model f_θ parameterized by θ. The model is first trained for the target segmentation task and adapted to output a per-pixel class-probability tensor $f_\theta(x) = \mathbf{p}_{\theta|x} \in [0, 1]^{C\times H\times W}$. It is then fine-tuned with EDL to produce Dirichlet predictive distributions, from which we derive pixel-wise uncertainty scores $s(u)$ and calibrate class-conditional thresholds.

At run-time, the calibrated thresholds $\tau_c(\alpha)$ are used for risk-controlled selective prediction. For each pixel u, the predicted label $\hat{y}(u)$ is accepted if its uncertainty score satisfies

$$s(u) \leq \tau_{\hat{y}(u)}(\alpha).$$

The core advantage of this design is that uncertainty estimation and risk calibration are defined independently of the internal segmentation architecture. This has two consequences: first, it enables a controlled comparison across different segmentation models under the same uncertainty and conformal protocol; and second, it supports inference with a uniform class-conditional accept/reject rule that can be recalibrated for a new dataset without redesigning the underlying segmentation backbone. Operationally, our design permits reuse of the same acceptance mechanism across datasets, with recalibration of thresholds rather than redesign of the model-specific uncertainty head.

3.1 Design-Time Process

The design-time process is a two-step procedure that is based on a given black-box segmentation model f_θ and a labeled dataset $\mathcal{D}$. We now detail the uncertainty-aware prediction (step one) and the conformal calibration (step two) of the design-time process.

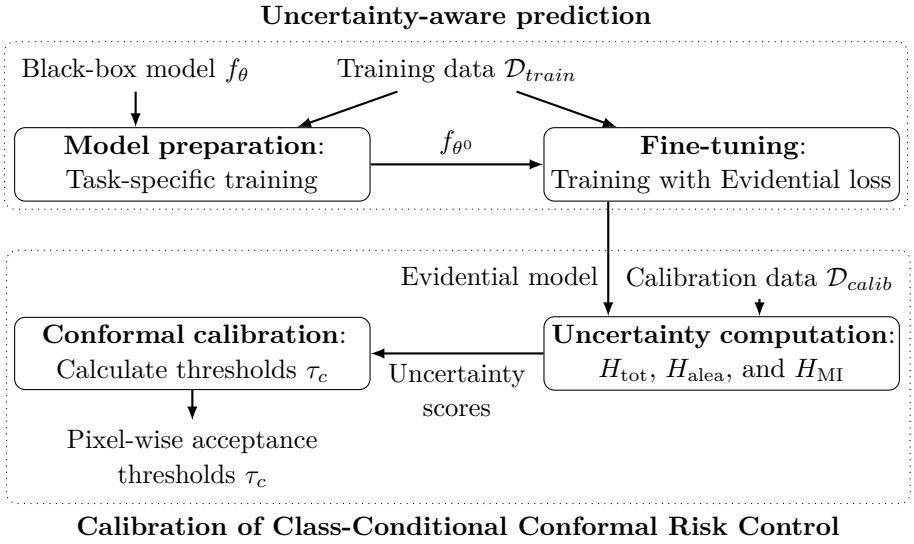

Figure 1 Designtime process for a model-agnostic uncertainty-aware semantic segmentation pipeline with risk-controlled selective prediction.

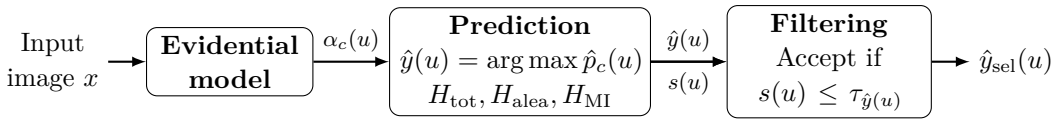

Figure 2 Runtime process: Conformal risk-control using class-conditional threholds τ_c to accept/reject class predictions per pixel with $\hat{y}_{\text{sel}}(u) \in \{1, \ldots, C, \emptyset\}$.

3.1.1 Uncertainty-Aware Prediction

The goal of this first step is to prepare the given segmentation model f_θ for uncertainty-aware prediction. Assume that the initial segmentation model outputs logits $z_\theta(x) \in \mathbb{R}^{C_\theta \times H \times W}$ for an input image is $x \in \mathbb{R}^{H \times W \times 3}$, where C_θ is the number of output channels of the original model and the last dimension of size 3 corresponds to the RGB color channels. We adapt the output layer so that the model matches the task-specific label space with C semantic classes, yielding $z'_\theta(x) \in \mathbb{R}^{C \times H \times W}$.

After applying softmax, we obtain the per-pixel class probabilities as expressed in Eq. 1

$$p_{\theta,c}(u) = \frac{\exp(z'_{\theta,c}(u))}{\sum_{k=1}^{C} \exp(z'_{\theta,k}(u))}, \tag{1}$$

where $p_{\theta,c}(u)$ denotes the predicted probability of class c at pixel u. The predicted label is then given as

$$\hat{y}(u) = \arg \max_{c=1,\ldots,C} p_{\theta,c}(u). \tag{2}$$

We split the available labeled data into training, calibration, and test sets, denoted by $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{cal}}$, and $\mathcal{D}_{\text{test}}$, respectively. The baseline parameters after task-specific training are denoted by θ_0, and the subsequent uncertainty-aware fine-tuning stage updates these parameters to θ.

As per standard supervised learning, each sample in each data set associates an input image x with its corresponding ground-truth labels $y \in \{1, \ldots, C\}^{H \times W}$.

Model preparation optimizes the task-specific model f_{θ^0} on $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{N}$ with a standard cross-entropy loss (ℓ_{CE}). We apply softmax to the model outputs to obtain per-pixel class probability $p_{\theta,c}(u)$ and calculate the standard cross-entropy with regularization:

$$\mathcal{L}_{\text{base}}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \underbrace{\left(-\frac{1}{|\Omega|}\sum_{u\in\Omega} \log p_{\theta,y_i(u)}(u)\right)}_{\ell_{\text{CE}}(x_i,y_i;\theta)} + \lambda_{\text{reg}} R(\theta), \tag{3}$$

In Eq. 3, $R(\theta)$ denotes a regularization term, for example, weight decay, $\lambda_{\text{reg}} \geq 0$ controls its strength, and Ω denotes the set of pixel locations. This term penalizes overly large parameter values and helps stabilize training.

Minimizing Eq. (3) yields the baseline model f_{θ^0}.

Evidential Dirichlet Fine-tuning starts from f_{θ^0} and augments the objective to encourage informative uncertainty estimates. Following the concept introduced in [2], we define the fine-tuning objective that combines the segmentation loss, an uncertainty-driven term ($\mathcal{L}_{\text{uq}}$), and a catastrophic-forgetting penalty:

$$L_{\text{fine}}(\theta) = \frac{1}{N_{\text{ft}}}\sum_{i=1}^{N_{\text{ft}}} \ell_{\text{CE}}(x_i, y_i; \theta) + \lambda_{\text{uq}}\mathcal{L}_{\text{uq}}(\theta) + \lambda_{\text{cf}}\mathcal{R}_{\text{cf}}(\theta, \theta^0), \tag{4}$$

where N_{ft} denotes the number of samples used in the uncertainty-aware fine-tuning stage.

Here, the standard cross-entropy loss is used to stabilize training and preserve the segmentation features learned by the baseline model. λ_{cf} is the catastrophic-forgetting regularization weight, while $\mathcal{R}_{\text{cf}}$ is the penalty term preserving shared parameters near baseline values.

Together with $\lambda_{\text{cf}} \geq 0$ and $\mathcal{R}_{\text{cf}}(\theta, \theta^0)$, the model is discouraged from drifting too far from the pre-trained model, thereby reducing catastrophic forgetting (see eq. 9). The uncertainty-driven term can be expressed as, $\mathcal{L}_{\text{uq}} = \frac{1}{|\Omega|}\sum_{u\in\Omega} \ell_{\text{EDL}}(u; e)$, driven from the Dirichlet properties. The uncertainty estimation is learned by the uncertainty-driven term $\mathcal{L}_{\text{uq}}(\theta)$, which is based on evidential deep learning (EDL) [1, 9].

Its core idea is that the model learns to output evidence values for each class, which are then used to parameterize a Dirichlet distribution $Dir(\alpha_0, \ldots, \alpha_C)$ over class probabilities. The Dirichlet distribution is a probability distribution over class-probability vectors on the simplex, parameterized by positive concentration parameters $\alpha_1(u), \ldots, \alpha_C(u)$.

For each pixel u, the logits $z_{\theta,c}(u)$ outputted by the model are mapped to non-negative evidence values using the $softplus + 1$ function. Summing over them produces the Dirichlet strength $S(u)$ at pixel u:

$$\alpha_c(u) = \text{softplus}(z_{\theta,c}(u)) + 1, \qquad S(u) = \sum_{k=1}^{C} \alpha_k(u), \tag{5}$$

Here, $\text{softplus}(t) = \log(1 + e^t)$, which ensures non-negative evidence values. The mean ($\hat{p}_c(u)$) and variance ($\text{Var}[\hat{p}_c(u)]$) of the Dirichlet distribution can be calculated and used to obtain the predicted class ($\hat{y}(u)$):

$$\hat{p}_c(u) = \frac{\alpha_c(u)}{S(u)}, \qquad \text{Var}[\hat{p}_c(u)] = \frac{\alpha_c(u)\big(S(u) - \alpha_c(u)\big)}{S(u)^2\big(S(u)+1\big)}, \qquad \hat{y}(u) = \underset{c=1,\ldots,C}{\text{argmax}}\, \hat{p}_c(u) \tag{6}$$

The Dirichlet distribution is interpreted as follows. The model's output logits are seen as evidence for their associated classes. High values indicate that the model has strong evidence for a certain class, which translates to concentrating probability mass on that class. Conversely, when the model has low evidence for all classes, it provides a low concentration, which is equivalent to a uniform distribution represented by the Dirichlet distribution. This indicates high uncertainty - and is the reasoning behind the composition of the EDL loss [1]:

$$\ell_{\mathrm{EDL}}(u;\, e) = \underbrace{\sum_{c=1}^{C} \big(y_c(u) - \hat{p}_c(u)\big)^2}_{\ell_{\mathrm{MSE}}(u)} + \underbrace{\sum_{c=1}^{C} \mathrm{Var}[\hat{p}_c(u)]}_{\text{variance penalty}} + \underbrace{\beta(e)\, \mathrm{KL}(\mathrm{Dir}(\tilde{\boldsymbol{\alpha}}(u)) \,\|\, \mathrm{Dir}(\mathbf{1}))}_{\mathcal{L}_{\mathrm{KL}}(u)}. \tag{7}$$

Here, the third term is a KL-divergence regularizer that penalizes evidence assigned to incorrect classes. To avoid suppressing evidence for the correct class, we apply a masking operation $\tilde{\alpha}_c(u) = y_c(u) + (1 - y_c(u))\, \alpha_c(u)$ and regularize the non-masked parameters toward the uniform Dirichlet prior $\mathrm{Dir}(\mathbf{1})$. The model should fit a uniform distribution, representing high uncertainty for evidence that does not belong to the ground-truth class. The regularization is annealed during training with weight $\beta(e) = \min\big(1, \frac{e}{10}\big)$ where e is the current training epoch.

The $\ell_{\mathrm{MSE}}(u)$ sums the squared deviations over all classes at pixel u. It is the per-pixel mean-squared error between the one-hot ground-truth label $\mathbf{y}(u)$ and the Dirichlet mean prediction $\hat{\mathbf{p}}(u)$. The error encourages the Dirichlet mean to match the target class distribution, as expressed in Eq. 8.

$$\ell_{\mathrm{MSE}}(u) = \sum_{c=1}^{C} \big(y_c(u) - \hat{p}_c(u)\big)^2 \tag{8}$$

Mitigating catastrophic forgetting is central to avoiding performance degradation when using a pre-trained black-box segmentation model for uncertainty-aware segmentation. We address two effects. Firstly, performance can degrade when all layers are updated simultaneously. Thus, we use staged freezing and progressive unfreezing, where early layers remain frozen for an initial set of epochs and task-specific components are updated first. Deeper layers are then gradually unfrozen, and the optimizer is reinitialized after each unfreezing step to match the active parameter set.

Secondly, we regularize toward the baseline parameters. Let Θ_{shared} be the set of parameters shared by the baseline and fine-tuned models. We apply an ℓ_2 penalty toward θ^0 as

$$\mathcal{R}_{\mathrm{cf}}(\theta, \theta^0) = \sum_{j \in \Theta_{\mathrm{shared}}} \big\|\theta_j - \theta_j^0\big\|_2^2. \tag{9}$$

The catastrophic-forgetting regularizer ($\mathcal{R}_{\mathrm{cf}}(\theta, \theta^0)$) is applied only to ($\Theta_{\mathrm{shared}}$). This anchors shared feature/segmentation parameters close to (θ^0) via Eq. (9) while leaving non-shared (e.g., uncertainty-specific) parameters unconstrained. Together with progressive unfreezing, this limits drift from the baseline representation while still enabling uncertainty-aware adaptation during fine-tuning.

3.1.2 Calibration of Class-Conditional Conformal Risk Control

The second step of the design-time process extracts finite-sample, distribution-free error control bounds by applying split-conformal calibration on top of the established pixel-wise uncertainty scores. From the predicted Dirichlet distribution, we derive three scalar uncertainty scores used for filtering and conformal calibration.

Firstly, we use the predictive entropy of the Dirichlet mean as a scalar measure of total predictive uncertainty, expressed as $H_{\text{tot}}(u) = -\sum_{c=1}^{C} \hat{p}_c(u) \log \hat{p}_c(u),$. This choice is natural because the Dirichlet mean $\hat{p}(u)$ is the predictive class distribution used for inference, and its entropy summarizes how concentrated or diffuse that distribution is. Low entropy indicates a confident prediction, whereas high entropy indicates ambiguity among classes.

To quantify aleatoric uncertainty, we use the expected predictive entropy under the Dirichlet distribution as expressed in Eq. 10

$$H_{\text{alea}}(u) = \mathbb{E}_{p \sim \text{Dir}(\alpha(u))}[H(p)]. \tag{10}$$

For a Dirichlet distribution with parameters $\alpha_1(u), \ldots, \alpha_C(u)$ and strength $S(u) = \sum_{c=1}^{C} \alpha_c(u)$, this quantity admits the closed-form expression, as defined in Eq. 11

$$H_{\text{alea}}(u) = \psi(S(u) + 1) - \sum_{c=1}^{C} \frac{\alpha_c(u)}{S(u)} \psi(\alpha_c(u) + 1), \tag{11}$$

where $\psi(\cdot)$ denotes the digamma function. This quantity is available in closed form from the Dirichlet parameters and therefore does not require sampling-based approximation. Its computation involves a sum over the (C) classes and thus has a linear cost per pixel in (C).

Finally, we define epistemic uncertainty as the mutual-information gap, expressed as

$$MI(u) = H_{\text{tot}}(u) - H_{\text{alea}}(u). \tag{12}$$

This decomposition separates total predictive uncertainty into a data-related component, captured by $H_{\text{alea}}(u)$, and a model-related component, captured by $MI(u)$.

All uncertainty scores are computed analytically from the Dirichlet parameters at each pixel. Their cost scales linearly with the number of classes and is negligible relative to the segmentation forward pass.

Furthermore, we use class-conditional split-conformal calibration. For each calibration pixel u, let $\hat{y}(u) = \arg\max_c \hat{p}_c(u)$ be the predicted class and let $s(u)$ be the selected uncertainty score. For a fixed class c, we stratify the calibration pixels predicted as class c and form $\mathcal{S}_c = \{(s(u), \mathbb{I}(\hat{y}(u) \neq y(u))) \mid \hat{y}(u) = c,\ (x, y) \in \mathcal{D}_{\text{cal}}\}$. Thus, $\mathcal{S}_c$ contains, for class c, the uncertainty score of each calibration pixel predicted as class c together with its error indicator. Here, $\mathbb{I}(\cdot)$ denotes the indicator function, which equals 1 if its argument is true and 0 otherwise.

This grouping is based on the predicted class rather than the true class because, at inference time, the true label is unavailable and the acceptance threshold must be selected from $\hat{y}(u)$ alone. We therefore calibrate one threshold per predicted class.

Given a threshold τ, we retain only pixels in $\mathcal{S}_c$ whose uncertainty score satisfies $s \leq \tau$. The empirical conditional risk among these retained pixels is defined as expressed in Eq. 13

$$\hat{R}_c(\tau) = \frac{\sum_{(s,\text{err}) \in \mathcal{S}_c} \text{err}\, \mathbb{I}(s \leq \tau)}{\max\left(1, \sum_{(s,\text{err}) \in \mathcal{S}_c} \mathbb{I}(s \leq \tau)\right)}. \tag{13}$$

where $\text{err} \in \{0, 1\}$ denotes the misclassification indicator for the corresponding calibration pixel.

For a target risk level $\alpha \in (0, 1)$, we then select the largest threshold whose empirical conditional risk does not exceed α:

$$\tau_c(\alpha) = \max\left\{\tau \in \mathbb{R} \;\middle|\; \hat{R}_c(\tau) \leq \alpha\right\}. \tag{14}$$

The resulting thresholds $\tau_c(\alpha)$ define the class-conditional acceptance rule used at inference time.

3.2 Run-Time Process

During deployment, the evidential model generates a predicted class $\hat{y}(u)$ and an uncertainty score $s(u)$ for each pixel u. A prediction is accepted only if its uncertainty score satisfies the calibrated class-conditional threshold, as $s(u) \leq \tau_{\hat{y}(u)}(\alpha)$. Otherwise, the prediction is rejected, and the pixel is labeled as uncertain; that is, no semantic class is emitted for that pixel.

Formally, the selective output is expressed as

$$\hat{y}^{\text{sel}}(u) = \begin{cases} \hat{y}(u), & \text{if } s(u) \leq \tau_{\hat{y}(u)}(\alpha), \\ \emptyset, & \text{otherwise}, \end{cases} \tag{15}$$

where $\emptyset$ denotes abstention on potentially unreliable predictions.

These decisions are naturally parameterized by the associated predicted class $\hat{y}(u)$, yielding a directly deployable acceptance rule. We evaluate coverage and risk on accepted pixels as

$$\begin{aligned} \text{coverage} &= \frac{|\{u \in \Omega_{\text{test}} : \ s(u) \leq \tau_{\hat{y}(u)}(\alpha)\}|}{|\Omega_{\text{test}}|}, \\ \text{risk} &= \frac{|\{u \in \Omega_{\text{test}} : \ s(u) \leq \tau_{\hat{y}(u)}(\alpha), \ \hat{y}(u) \neq y(u)\}|}{|\{u \in \Omega_{\text{test}} : \ s(u) \leq \tau_{\hat{y}(u)}(\alpha)\}|}. \end{aligned} \tag{16}$$

Under exchangeability between calibration and test samples, the split-conformal calibration step in Eq. (14) provides a finite-sample guarantee for the conformal selection rule. In particular, the accepted set is calibrated so that its error rate is controlled in expectation compared to the target level α, up to sampling variability.

4 Evaluation

This section evaluates predictive performance, uncertainty behavior, and conformal risk control behavior for *heterogeneous transformer-based segmentation backbones*, comprising DINOv2 version *ViT-Base*, Mask2Former version *SwinV2-L*, and SegFormer version *MiT-B5*. These architectures are well-known for providing complementary inductive biases and represent widely adopted design paradigms for evaluating the effects of foundation pretraining, mask-based decoding, and efficiency-oriented semantic decoding across diverse domains.

We decompose predictive uncertainty into aleatoric and epistemic components and assess how these quantities support selective filtering and conformal risk control. Aleatoric uncertainty is quantified by expected predictive entropy and reflects data-related ambiguity, whereas epistemic uncertainty is measured by mutual information and reflects model uncertainty. The following evaluations are considered:

- **(A) Per-class aleatoric conformal thresholds**: class-conditional thresholds on aleatoric entropy as a function of target error level α. Decreasing thresholds as α increases indicate stricter acceptance, while class separation reflects class-dependent filtering.
- **(B) Per-class epistemic conformal thresholds**: analogous thresholds for epistemic MI, showing how model-uncertainty gating adapts across classes.
- **(C) Aleatoric entropy vs. error**: distributions of aleatoric entropy for correct and incorrect pixels. Stronger separation indicates better error discrimination.
- **(D) Epistemic MI vs. error**: distributions of epistemic MI for correct and incorrect pixels. Right-shifted error distributions support epistemic gating.
- **(E) Combine rule**: logical rule used to combine aleatoric and epistemic acceptance conditions; `and` accepts a pixel only if both scores are below their class-specific thresholds.

Moreover, results report standard segmentation metrics (accuracy, IoU, precision, recall, F1, AP, FW_IoU) [14], calibration metrics (pixel accuracy, ECE, AURC) [6], uncertainty statistics (correct/wrong-pixel means for total, aleatoric, and epistemic uncertainty) [11], and class-conditional conformal metrics (coverage, risk, aleatoric/epistemic thresholds per class at varying α) [15].

We evaluate our method primarily on the newly collected LiSS dataset. Since LiSS is not yet publicly available, we additionally report cross-dataset experiments on the COCO dataset [1] in Appendix A to support reproducibility and facilitate comparison on a widely used benchmark.

4.1 Experimental Setup

To support understanding of the urban scene conditions, we mimic realistic operating conditions by collecting first-hand video data across multiple districts of Lisbon, Portugal, using bicycle-mounted recording devices. Data acquisition is performed at different times of day to capture variability in illumination, weather, and crowd density. The resulting high-definition video streams are temporally resampled at 30 frames per second to obtain image frames suitable for semantic segmentation while limiting redundancy between adjacent frames

Using the obtained image frames, segmentation masks are generated through an automated, modular preprocessing pipeline. Object instances are detected using YOLOv8 to obtain bounding-box proposals, which are then used as prompts for SAM to produce instance-level masks aligned with object boundaries. A subsequent HSV-based image-processing stage is applied to refine the separation between *path* and non-path regions via thresholding and morphological filtering. This improves class consistency in visually homogeneous areas. The final annotations are encoded into three classes: *path* (0), *object* (1), and *background* (2). After manual visual validation on a subset of samples, the dataset comprises $171,098$ image-mask pairs and is split into training (70%), validation (15%), and test (15%) sets with sequence-level separation to prevent leakage.

The overall model configurations, training setup, and conformal calibration parameters used with evidential fine-tuning are summarized in Table 9 in Appendix B.

4.2 In-Domain Results: LiSS Dataset

We first analyze model behavior on the LiSS dataset to assess in-domain segmentation performance, uncertainty-error alignment, and class-conditional risk control under the proposed pipeline. Table 1 reports global performance on the LiSS dataset. Mask2Former achieves the highest pixel accuracy (0.9726) and the lowest AURC (0.0025). It also has the lowest ECE (0.05203), indicating better calibration compared to DINOv2 and SegFormer. However, its epistemic MI separation is small (0.2218 correct vs. 0.2235 wrong), so epistemic uncertainty provides limited error discrimination. DINOv2 attains an accuracy of (0.9687) with ECE (0.0863). It shows clear uncertainty gaps for aleatoric entropy ($0.3073 \rightarrow 0.7702$) and epistemic MI ($0.0336 \rightarrow 0.0776$), which supports uncertainty-based filtering. **SegFormer** has the lowest accuracy (0.9596) and the highest ECE (0.0979). It achieves a low AURC (0.0028) and shows strong uncertainty gaps (aleatoric: $0.3782 \rightarrow 0.8009$, epistemic: $0.0420 \rightarrow 0.0870$), which also supports uncertainty-based filtering.

[1] `https://cocodataset.org/#home`

Table 1 Global performance and error-conditioned uncertainty. Pixel-wise accuracy, expected calibration error (ECE), and selective prediction quality (AURC) for DINOv2, Mask2Former, and SegFormer on the ***Lisbon street scene dataset*** *`"alea"` and `"epi"` stand for aleatoric and epistemic, respectively.*

Model	Metric								
	pixel accuracy	ece	aurc	total correct	total wrong	alea correct	alea wrong	epi correct	epi wrong
DINOv2	0.9687	0.0863	0.0030	0.3410	0.8478	0.3073	0.7702	0.0336	0.0776
Mask2Former	0.9726	0.05203	0.0025	1.0670	1.0801	0.8452	0.8608	0.2218	0.2235
SegFormer	0.9596	0.0979	0.0028	0.4202	0.8880	0.3782	0.8009	0.0420	0.0870

Table 2 Class-wise segmentation performance for DINOv2, Mask2Former, and SegFormer across the evaluated semantic classes on the ***Lisbon street scene dataset***.

Model	Class	Metric							
		accuracy	*IoU*	*MPA*	*precision*	*recall*	*F1*	*AP*	*FW_IoU*
DINOv2	***path***	0.9838	0.9463	0.9955	0.9503	0.9955	0.9724	0.9970	0.9408
	obstacle	0.9839	0.8354	0.9907	0.8421	0.9907	0.9103	0.9789	0.9408
	background	0.9698	0.9521	0.9536	0.9984	0.9536	0.9755	0.9980	0.9408
Mask2Former	***path***	0.9916	0.9763	0.9941	0.9820	0.9941	0.9880	0.9978	0.9501
	object	0.9796	0.6692	0.8959	0.7257	0.8959	0.8018	0.9855	0.9501
	background	0.9734	0.9565	0.9656	0.9902	0.9656	0.9777	0.9991	0.9501
SegFormer	***path***	0.9792	0.5671	0.5929	0.9289	0.5929	0.7238	0.9579	0.9376
	object	0.9682	0.9497	0.9926	0.9565	0.9926	0.9742	0.9978	0.9376
	background	0.9879	0.9655	0.9736	0.9914	0.9736	0.9824	0.9958	0.9376

4.2.1 Class-Wise Segmentation Quality

We next examine class-wise segmentation metrics to understand how each model distributes its successes and failures across the different semantic classes. Table 2 reveals complementary class-wise strengths of each model. The table shows that Mask2Former obtains the highest FW_IoU (0.9501) and strongest *path* IoU (0.9763), consistent with mask-based decoding favoring coherent large regions, but shows weaker *object* performance (0.6692 on IoU). SegFormer achieves the highest *object* IoU (0.9497) and strong *background* IoU (0.9655), but its *path* IoU (0.5671) is significantly reduced by low recall (0.5929), indicating under-segmentation of traversable regions. DINOv2 provides the most balanced behavior across classes, with strong *path* (0.9463) and *background* (0.9521) IoU and moderate *obstacle* performance (0.8354), reducing extreme class-specific failure modes.

4.2.2 Uncertainty Characteristics and Conformal Thresholds

Figures 3 and 4 show class-conditional conformal thresholds under the *and* rule. For aleatoric thresholds (Figure 3), DINOv2 shows the clearest class dependence, with *path* threshold remaining high as α increases (e.g., $0.8266 \rightarrow 0.9533$) and *obstacle* threshold starting much lower. Mask2Former is comparatively uniform, with weak class separation and limited variation across α. SegFormer, on the other hand, shows stronger increases in the class-conditional conformal thresholds for *path* and *obstacle* as α increases, while the *background* threshold stays consistently high.

Figure 4 indicates that, for SegFormer, uncertainty-based filtering is less effective at rejecting error-prone pixels at low target risk levels. Even when many pixels are rejected (Table 3), the conditional error on the remaining (accepted) pixels can stay large at low (α). Under the *and* rule, decreasing the epistemic threshold can only increase (or leave unchanged)

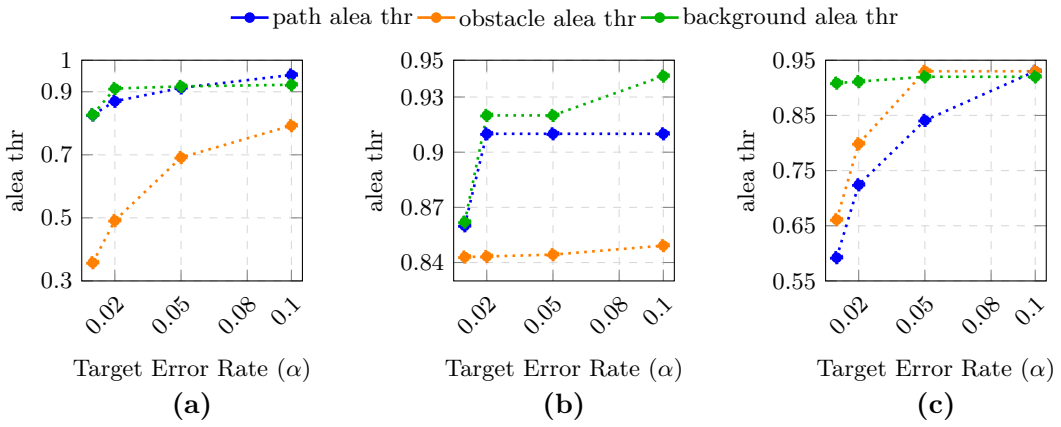

Figure 3 Class-conditional conformal thresholds on aleatoric entropy versus target error level α for **(a) DINOv2**, **(b) Mask2Former**, and **(c) SegFormer** on the Lisbon street scene dataset. Curves summarize how strictly each model filters predictions based on data-driven uncertainty across classes. (***"alea thr"*** is the short version for aleatoric threshold).

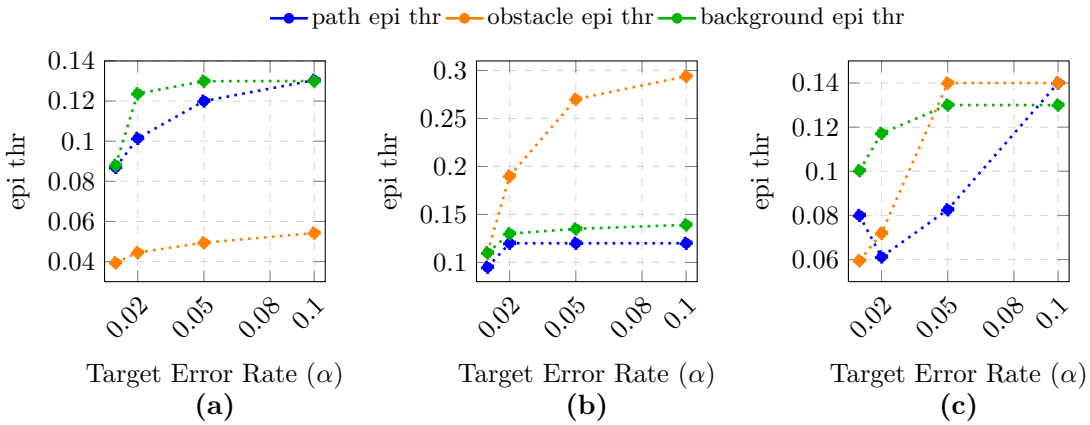

Figure 4 Class-conditional conformal thresholds on epistemic entropy (mutual information) versus target error level α for **(a) DINOv2**, **(b) Mask2Former**, and **(c) SegFormer** on the Lisbon street scene dataset. Curves show how model-uncertainty thresholds adapt across classes for different error tolerances. (***"epi thr"*** is the short version for epistemic threshold).

the number of rejected pixels. However, the conditional risk on the accepted set may still increase at low (α) if the extra rejections are dominated by `correct but uncertain` pixels rather than errors. In contrast, DINOv2 (and, to a lesser extent, Mask2Former) shows improved epistemic thresholding, rejecting a larger fraction of error-prone pixels, leading to more stable or reduced conditional risk in this low-(α) range. This behavior is reflected in the rejected-pixel counts and corresponding risks in Table 3.

Figures 5 and 6 compare error-conditioned uncertainty distributions across the three models. Considering the aleatoric entropy (Figure 5), DINOv2 and SegFormer show a clear right shift for wrong predictions compared to correct predictions. This indicates greater uncertainty in the errors. Mask2Former concentrates in a narrow, high-entropy region and shows weaker separation between correct and incorrect predictions, limiting its discriminative ability for filtering. This can be attributed to the evidential/Dirichlet outputs yielding similar

probability 'softness' across pixels, so entropy doesn't vary much with errors. Looking at the epistemic MI (Figure 6), DINOv2 and SegFormer again exhibit higher MI for wrong predictions, consistent with meaningful epistemic separation. Mask2Former, on the other hand, remains highly compressed in MI with minimal separation, suggesting weaker epistemic error discrimination.

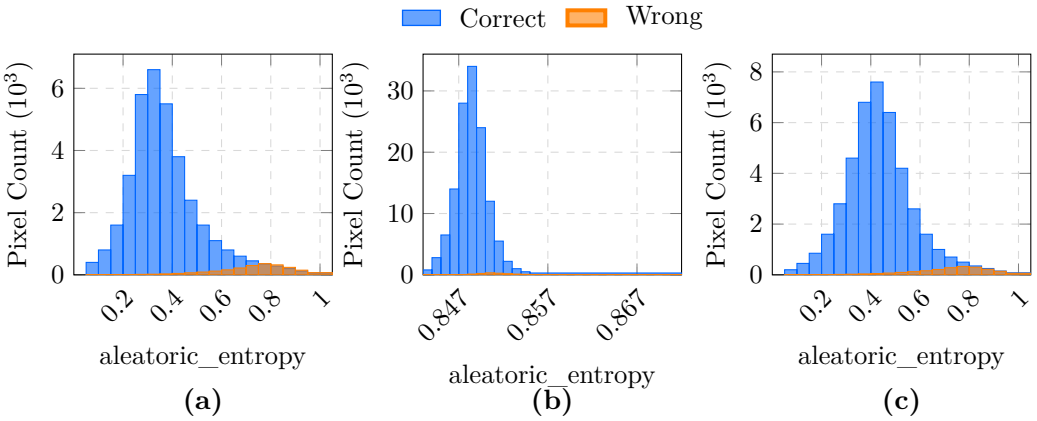

Figure 5 Aleatoric entropy vs. error. Distribution of aleatoric entropy for correct and wrong predictions across sampled pixels for **(a) DINOv2**, **(b) Mask2Former**, and **(c) SegFormer** on the Lisbon street scene dataset. **The y-axis shows pixel counts in thousands.**

4.2.3 Class-Conditional Risk Control

Table 3 Risk and safety performance evaluation. Class-conditional coverage, risk, and conformal thresholds for aleatoric and epistemic uncertainty under the `"and"` combination rule for the DINOv2, Mask2Former, and SegFormer models on the ***Lisbon street scene dataset***.

Model	α	Metric						
		combine rule	*path coverage*	*path risk*	*path alea thr*	*path epi thr*	*Path rejected pixels*	*obstacle coverage*	*obstacle risk*
DINOv2	***0.01***	***and***	0.9757	0.0050	0.8266	0.0869	83524093	0.6721	0.3739
	0.02	***and***	0.9939	0.0137	0.8698	0.1014	20839755	0.7140	0.4228
	0.05	***and***	0.9983	0.0135	0.9123	0.1200	5713302	0.7330	0.3659
	0.10	***and***	0.9983	0.0139	0.9533	0.1304	5710321	0.7506	0.3185
Mask2Former	***0.01***	***and***	0.9339	0.0125	0.8600	0.0950	227031915	0.9329	0.6500
	0.02	***and***	0.9997	0.0136	0.9100	0.1200	961691	0.9392	0.9000
	0.05	***and***	0.9997	0.0136	0.9100	0.1200	961691	0.9392	0.9000
	0.10	***and***	0.9997	0.0136	0.9100	0.1200	961691	0.9425	0.9000
SegFormer	***0.01***	***and***	0.9917	0.9800	0.5922	0.0800	28389482	0.8908	0.0056
	0.02	***and***	0.9934	0.4187	0.7242	0.0612	22648206	0.9375	0.0054
	0.05	***and***	0.9982	0.4419	0.8409	0.0826	6076860	0.9980	0.0065
	0.10	***and***	0.9983	0.3535	0.9300	0.1400	6004210	0.9980	0.0089

Tables 3 and 4 report selective prediction under the `"and"` rule, where a pixel is accepted only if both aleatoric and epistemic scores fall below the class-specific thresholds.

DINOv2 maintains high *path* coverage (e.g., 0.9757 at α=0.01 and $\approx$ 0.9983 for larger α) and high *background* coverage (e.g., 0.9622 at α=0.01). Its *path* risk remains low (on the order of 10^{-2}), while *obstacle* coverage is noticeably lower than the other classes (roughly 0.67–0.75), with comparatively higher *obstacle* risk.

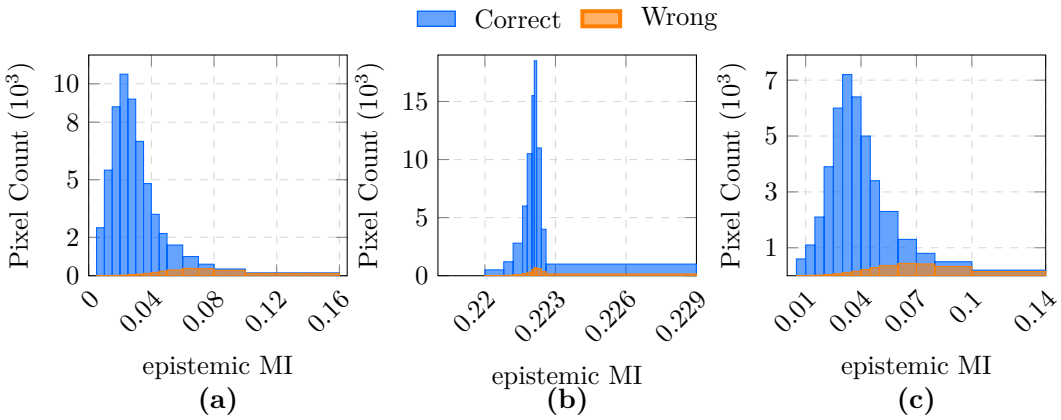

Figure 6 Epistemic MI vs. error. Distribution of epistemic uncertainty (mutual information) for correct and wrong predictions across sampled pixels for **(a) DINOv2**, **(b) Mask2Former (zoomed view)**, and **(c) SegFormer** on the Lisbon street scene dataset. **The y-axis shows pixel counts in thousands.**

Table 4 Continuation of Table 3.

Model	α	Metric								
		combine rule	*obstacle alea thr*	*obstacle epi thr*	*obstacle rejected pixel*	*bg coverage*	*bg risk*	*bg alea thr*	*bg epi thr*	bg rejected pixel
DINOv2	***0.01***	***and***	0.3578	0.0394	148482680	0.9622	0.0038	0.8272	0.0880	226745123
	0.02	***and***	0.4914	0.0445	129532145	0.9890	0.0107	0.9101	0.1237	65688610
	0.05	***and***	0.6915	0.0494	120912361	0.9898	0.0127	0.9170	0.1299	61271376
	0.10	***and***	0.7926	0.0542	112923523	0.9899	0.0129	0.9222	0.1300	60832283
Mask2Former	***0.01***	***and***	0.8430	0.1100	30371812	0.9361	0.0105	0.8620	0.1100	383014065
	0.02	***and***	0.8433	0.1900	27551165	0.9983	0.0524	0.9200	0.1300	10242614
	0.05	***and***	0.8443	0.2701	27551165	0.9983	0.0588	0.9200	0.1350	10242614
	0.10	***and***	0.8492	0.2940	26031046	0.9983	0.0595	0.9414	0.1392	10241006
SegFormer	***0.01***	***and***	0.6609	0.0595	49443364	0.9985	0.0033	0.9086	0.1003	8850235
	0.02	***and***	0.7987	0.0719	28317145	0.9987	0.0081	0.9113	0.1172	7802399
	0.05	***and***	0.9300	0.1400	916905	0.9999	0.0413	0.9200	0.1300	202239
	0.10	***and***	0.9300	0.1400	916905	0.9999	0.0413	0.9200	0.1300	202239

Mask2Former shows a clear transition from more conservative acceptance at α=0.01 (e.g., *path* coverage 0.9339 and *background* coverage 0.9361) to near-complete acceptance at larger α (both *path* and *background* ≈ 0.9983). However, the reported risks for *obstacle* and *background* increase with α, suggesting that relaxing the thresholds admits more difficult instances and highlighting a trade-off between coverage and reliability for these classes.

SegFormer achieves very high *background* coverage (approaching 0.9999 at $\alpha \geq 0.05$) and increasing *obstacle* coverage (from 0.8908 at α=0.01 to ≈ 0.9980 for larger α). Its *obstacle* risk remains low (below 10^{-2}), but *path* risk is consistently high (e.g., 0.9800 at α=0.01 and 0.3535 at α=0.10). This pattern indicates that SegFormer failures on *path* are not well captured by the filtering-based uncertainty scores, even when acceptance is high.

4.3 Discussion

These results show that reliable risk-controlled segmentation depends on the interaction between backbone design, uncertainty estimation, and conformal calibration. In-domain on LiSS, all models achieve high pixel accuracy and low AURC, but they differ in how

informative their uncertainties are. DINOv2 and SegFormer exhibit clear correct-to-wrong gaps in both aleatoric entropy and epistemic MI, which supports uncertainty-based filtering, while Mask2Former shows almost no epistemic MI separation despite strong global calibration metrics.

Class-conditional conformal evaluation on LiSS further highlights model-dependent behavior. DINOv2 achieves low path risk with high coverage, whereas SegFormer exhibits persistently high path risk even at high coverage, suggesting that its failures on this class are not well captured by the filtering-based uncertainty scores. Mask2Former achieves stable coverage, but its MI remains compressed, limiting epistemic adaptivity across classes.

Under cross-dataset transfer to the COCO dataset, global metrics remain relatively strong, but class-wise performance is highly uneven (see Appendix A). SegFormer retains the strongest per-class IoU and FW_IoU on the selected COCO classes, while DINOv2 degrades substantially on vehicle classes, and Mask2Former collapses on background and person, despite isolated strengths on *"bus"* class.

Updated conformal results on COCO (Tables 7 and 8 in Appendix A) show improved class-conditional risk control compared to the earlier version. *"Background"* risks are low across models and α, and *"person"* risks increase moderately as coverage increases, reflecting a clearer coverage-risk trade-off and more consistent threshold adaptation.

For autonomous navigation, the practical implication is that risk control should be validated per class and per domain, rather than assumed from global accuracy or calibration alone. Our proposed pipeline provides a uniform interface for rejection, but deployment still requires monitoring of class-wise failure modes and periodic recalibration when the operating domain or label semantics change.

The risk guarantees and the thresholds provided by our work should be interpreted carefully in real-world applications. The formal component arises from the split-conformal calibration step and relies on exchangeability between calibration and test data. It does not imply that the uncertainty scores themselves are intrinsically calibrated or that the same risk control will hold unchanged under domain shift. The cross-dataset results on COCO illustrate this limitation, as weaker alignment between uncertainty and error reduces the effectiveness of the calibrated acceptance rule.

5 Conclusion and Future Work

This work investigates model-agnostic uncertainty-aware semantic segmentation with explicit risk control. We introduced a unified evaluation pipeline that treats segmentation architectures as black-box models producing per-pixel class probabilities. Each evaluated model is augmented with evidential Dirichlet fine-tuning to obtain aleatoric and epistemic uncertainty signals. We then apply pixel-wise, class-conditional conformal calibration to derive acceptance thresholds for user-specified target error levels. The results show that uncertainty-error alignment and the resulting selective filtering behavior depend on the underlying architecture and on the semantic class, and that these indicators are weaker under cross-dataset evaluation. This, in turn, reduces the effectiveness and adaptivity of uncertainty-based conformal filtering. Key limitations of this work suggest three main directions for future work: improving conformal calibration under dataset shift, exploring richer nonconformity and spatially structured uncertainty measures for better pixel-wise risk control, and developing more principled cross-dataset label alignment to reduce ambiguity and improve transfer.

References

1 Siddharth Ancha, P. R. Osteen, and Nicholas Roy. Deep evidential uncertainty estimation for semantic segmentation under out-of-distribution obstacles. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2024. URL: http://siddancha.github.io/projects/evidential-segmentation-uncertainty/paper.pdf.

2 Siddharth Ancha, Philip R Osteen, and Nicholas Roy. Deep evidential uncertainty estimation for semantic segmentation under out-of-distribution obstacles. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6943–6951. IEEE, 2024. doi:10.1109/ICRA57147.2024.10611342.

3 Di Chen, Ziwei Liu, Chen Yang, and Dong Wang. Conformalsam: Unlocking the potential of foundational segmentation models in semi-supervised semantic segmentation with conformal prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. URL: https://openaccess.thecvf.com/content/ICCV2025/html/Chen_ConformalSAM_Unlocking_the_Potential_of_Foundational_Segmentation_Models_in_Semi-Supervised_ICCV_2025_paper.html.

4 Bowen Cheng, Ishan Misra, Alexander G. Schwing, and Alexander Kirillov. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. URL: http://openaccess.thecvf.com/content/CVPR2022/html/Cheng_Masked-Attention_Mask_Transformer_for_Universal_Image_Segmentation_CVPR_2022_paper.html.

5 Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. URL: http://openaccess.thecvf.com/content_cvpr_2016/html/Cordts_The_Cityscapes_Dataset_CVPR_2016_paper.html.

6 Yifan Ding, Jiayang Liu, Jing Xiong, and Yanzhe Shi. Revisiting the evaluation of uncertainty estimation and its application to explore model complexity-uncertainty trade-off. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020. URL: http://openaccess.thecvf.com/content_CVPRW_2020/papers/w1/Ding_Revisiting_the_Evaluation_of_Uncertainty_Estimation_and_Its_Application_to_Explore_Model_Complexity-Uncertainty_Trade-off_CVPRW_2020_paper.pdf.

7 Jiarui Gao, Meng Chen, Lei Xiang, and Chen Xu. A comprehensive survey on evidential deep learning and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. Early access. URL: https://arxiv.org/abs/2409.04720.

8 Daniel-Valentin Giurgi, Mulugeta N. Geletu, Thomas Josso-Laurain, and Afef M. Ahrary. Intelligent vehicles semantic segmentation using evidential deep learning. In *Upper Rhine Artificial Intelligence Symposium (URAI)*, 2023. URL: https://hal.science/hal-04763422.

9 Yuyang He and Lei Li. Uncertainty-aware evidential fusion-based learning for semi-supervised medical image segmentation. *arXiv preprint arXiv:2404.06177*, 2024. doi:10.48550/arXiv.2404.06177.

10 Lior Hirschfeld, Kyle Swanson, Kevin Yang, Regina Barzilay, and Connor W Coley. Uncertainty quantification using neural networks for molecular property prediction. *Journal of Chemical Information and Modeling*, 60(8):3770–3780, 2020. doi:10.1021/ACS.JCIM.0C00502.

11 Sunghyun Kim and Cheol Soo Park. Quantification of aleatoric and epistemic uncertainty in data-driven occupant behavior model for building performance simulation. *Energy and Buildings*, page 115818, 2025.

12 Ioannis Konidakis, Kyriaki Panayidou, and Grigorios Tsagkatakis. Uncertainty-aware flood segmentation from Sentinel-2 observations with conformal prediction. In *IGARSS 2025 – IEEE International Geoscience and Remote Sensing Symposium*, 2025. URL: https://users.ics.forth.gr/~tsakalid/PAPERS/CNFRS/2025-IGARSS-Konidakis.pdf.

13 Shiman Li, Mingzhi Yuan, Xiaokun Dai, and Chenxi Zhang. Evaluation of uncertainty estimation methods in medical image segmentation: Exploring the usage of uncertainty in clinical deployment. *Computerized Medical Imaging and Graphics*, page 102574, 2025. doi:10.1016/J.COMPMEDIMAG.2025.102574.

14 Romain Martin and Long Duong. Pixel-wise confidence estimation for segmentation in bayesian convolutional neural networks. *Machine Vision and Applications*, 2023. doi:10.1007/s00138-022-01369-9.

15 Lucile Mossina, Joaquim Dalmau, and Lucas Andéol. Conformal semantic image segmentation: Post-hoc quantification of predictive uncertainty. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024. URL: https://openaccess.thecvf.com/content/CVPR2024W/SAIAD/html/Mossina_Conformal_Semantic_Image_Segmentation_Post-hoc_Quantification_of_Predictive_Uncertainty_CVPRW_2024_paper.html.

16 Maxime Oquab, Théo Darcet, Théodore Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre-Yves Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. doi:10.48550/arXiv.2304.07193.

17 Qingtao Pan, Zhen Li, Guang Yang, Qing Yang, and Bo Ji. Evivlm: When evidential learning meets vision language model for medical image segmentation. *IEEE Transactions on Medical Imaging*, 2025. Early access. URL: https://www.researchgate.net/publication/396514745.

18 Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint*, 2025. arXiv:2508.10104.

19 Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. URL: https://proceedings.neurips.cc/paper/2021/file/64f1f27bf1b4ec22924fd0acb550c235-Paper.pdf.

20 Fang Zhang, Jun Wang, Wei Fan, and Shuai Guo. Fast, trusted semantic segmentation of lidar point clouds with evidential fusion. In *Proceedings of SPIE, Conference on Graphics and Image Processing*, 2025. URL: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/13539/135390M.

A Cross-Dataset Transfer: COCO Evaluation

For the COCO evaluation, images and annotations are used directly from the benchmark. Only standard preprocessing, including resizing and normalization, is applied, with no additional YOLO, SAM, or HSV-based annotation stages. Instance masks are consolidated into class-wise semantic masks; when instances overlap, each pixel is assigned to the instance with the highest mask score, while pixels outside the selected classes are labeled as *background*. To support cross-domain comparison with LiSS, the COCO experiments are limited to three frequent and safety-relevant classes: *person*, *car*, and *bus*.

To assess generalization beyond LiSS, we evaluate all models on COCO using a restricted set of safety-relevant classes (*person*, *car*, *bus*, *background*). This comparison is important because LiSS is newly collected and not publicly available, whereas COCO provides a widely used benchmark supporting reproducible evaluation.

Table 5 Global performance and error-conditioned uncertainty on the COCO dataset. Reported metrics include pixel accuracy, ECE, AURC, and mean uncertainty values on correctly and incorrectly classified pixels.

Model	Metric								
	pixel accuracy	ece	aurc	total correct	total wrong	alea correct	alea wrong	epi correct	epi wrong
DINOv2	0.8350	0.0901	0.0065	0.3633	0.7827	0.3440	0.7200	0.0902	0.1823
Mask2Former	0.8720	0.0650	0.0050	0.3821	0.7611	0.3619	0.7000	0.1000	0.1902
SegFormer	0.8480	0.0852	0.0060	0.3709	0.7900	0.3501	0.7100	0.0950	0.1850

Table 6 Per-class segmentation performance on the COCO dataset. The table reports the class-wise average pixel-level classification performance across the three models.

Model	Class	Metric							
		accuracy	*IoU*	*MPA*	*precision*	*recall*	*F1*	*AP*	*FW_IoU*
DINOv2	*background*	0.4234	0.2872	0.2876	0.9949	0.2876	0.4463	0.9893	0.2725
	person	0.5221	0.2476	0.9783	0.2489	0.9783	0.3969	0.7669	0.2725
	car	0.9392	0.0697	0.4647	0.0758	0.4647	0.1303	0.0000	0.2725
	bus	0.9704	0.0005	0.0011	0.0010	0.0011	0.0011	0.0000	0.2725
Mask2Former	*background*	0.2005	0.0096	0.0096	1.0000	0.0096	0.0190	0.9983	0.0468
	person	0.3751	0.1989	0.9965	0.1991	0.9965	0.3318	0.9833	0.0468
	car	0.8959	0.0840	0.9106	0.0847	0.9106	0.1549	0.0000	0.0468
	bus	0.9854	0.4034	0.9437	0.4134	0.9437	0.5749	0.0000	0.0468
SegFormer	*background*	0.8175	0.7747	0.7767	0.9966	0.7767	0.8731	0.9993	0.7120
	person	0.8520	0.5183	0.9909	0.5208	0.9909	0.6827	0.9660	0.7120
	car	0.9784	0.2970	0.9314	0.3037	0.9314	0.4580	0.0000	0.7120
	bus	0.9732	0.0012	0.0023	0.0026	0.0023	0.0024	0.0000	0.7120

Table 7 Class-conditional conformal prediction on COCO. Per-class coverage, risk, and calibrated thresholds for aleatoric and epistemic uncertainty under the `"and"` combination rule.

Model	α	Metric									
		bg coverage	*bg risk*	*bg alea thr*	*bg epi thr*	*bg rejected pixels*	*person coverage*	*person risk*	*person alea thr*	*person epi thr*	*person rejected pixels*
DINOv2	***0.01***	0.852	0.012	0.721	0.090	47888495	0.250	0.080	0.600	0.082	44867675
	0.02	0.880	0.014	0.770	0.111	38310796	0.321	0.091	0.650	0.102	40680026
	0.05	0.920	0.018	0.821	0.152	25540531	0.451	0.111	0.731	0.141	32902962
	0.10	0.953	0.025	0.880	0.201	15962832	0.601	0.140	0.820	0.190	23929427
Mask2Former	***0.01***	0.881	0.011	0.752	0.104	38310796	0.302	0.072	0.623	0.091	41876497
	0.02	0.919	0.012	0.792	0.124	28733097	0.381	0.081	0.682	0.110	37090612
	0.05	0.941	0.016	0.843	0.161	19155398	0.524	0.101	0.761	0.154	28715312
	0.10	0.965	0.022	0.900	0.211	11173982	0.680	0.132	0.850	0.201	19143541
SegFormer	***0.01***	0.824	0.015	0.701	0.090	57466194	0.221	0.092	0.582	0.080	46662382
	0.02	0.863	0.017	0.741	0.111	44695929	0.301	0.101	0.641	0.102	41876497
	0.05	0.901	0.021	0.792	0.154	31925664	0.421	0.120	0.721	0.141	34697669
	0.10	0.931	0.028	0.853	0.200	22347965	0.561	0.152	0.810	0.191	26322369

Table 8 Continuation of Table 7.

Model	α	Metric									
		bus coverage	*bus risk*	*bus alea thr*	*bus epi thr*	*bus rejected pixels*	*car coverage*	*car risk*	*car alea thr*	*car epi thr*	*car rejected pixels*
DINOv2	***0.01***	0.020	0.163	0.550	0.074	4435824	0.034	0.140	0.571	0.075	4113343
	0.02	0.041	0.171	0.614	0.090	4345297	0.061	0.152	0.634	0.095	3986126
	0.05	0.072	0.193	0.710	0.139	4209506	0.112	0.176	0.721	0.135	3816504
	0.10	0.120	0.220	0.843	0.181	3983189	0.171	0.204	0.821	0.185	3519665
Mask2Former	***0.01***	0.031	0.152	0.588	0.081	4390560	0.040	0.132	0.601	0.085	4070938
	0.02	0.060	0.164	0.650	0.100	4254770	0.081	0.142	0.671	0.105	3901315
	0.05	0.101	0.180	0.740	0.149	4073716	0.133	0.164	0.761	0.145	3689287
	0.10	0.162	0.210	0.851	0.197	3802135	0.238	0.191	0.864	0.195	3392448
SegFormer	***0.01***	0.015	0.188	0.533	0.070	4458456	0.025	0.166	0.552	0.075	4134546
	0.02	0.030	0.190	0.590	0.091	4390560	0.050	0.171	0.610	0.095	4028532
	0.05	0.061	0.218	0.682	0.130	4254770	0.098	0.190	0.717	0.135	3858910
	0.10	0.118	0.241	0.781	0.180	4073716	0.150	0.224	0.800	0.186	3604476

Table 5 summarizes cross-dataset performance on COCO and contrasts it with the in-domain LiSS results in Table 1. All three models retain reasonably high pixel accuracy on COCO ($\approx$ 0.8350–0.8720), with Mask2Former achieving the highest accuracy ($\approx$ 0.8720)

and the lowest ECE (≈ 0.0650). Selective prediction degrades only mildly under shift, as AURC remains low (≈ 0.0050–0.0065) but increases relative to LiSS. The correct-to-wrong uncertainty gaps persist across all models, indicating that uncertainty still correlates with errors on COCO, although the separation is weaker than on the LiSS dataset.

A.1 Class-Dependent Transfer

Table 6 shows a strongly class-dependent transfer. SegFormer achieves the highest FW_IoU (0.7120) and strongest IoU on *background* (0.7747), *person* (0.5183), and *car* (0.2970). DINOv2 shows substantially lower IoU, particularly for *car* (0.0697) and nearly zero for *bus* (0.0005). Mask2Former attains very low FW_IoU (0.0468), driven by poor *background* (0.0096) and *person* (0.1989) IoU, despite comparatively higher *bus* IoU (0.4034). These results indicate that transfer to COCO is highly uneven across the selected classes, with SegFormer retaining the most robust cross-domain performance, while DINOv2 and especially Mask2Former exhibit severe degradation in key foreground classes (notably *car*) despite isolated strengths.

A.2 Conformal Risk Control

Tables 7 and 8 indicate substantially improved class-conditional risk control on COCO. For *background*, all models achieve low conditional risk across α (e.g., DINOv2: $0.0120 \rightarrow 0.0253$, Mask2Former: $0.0108 \rightarrow 0.0220$ as α increases from 0.01 to 0.10), while coverage increases and the number of rejected pixels decreases. This behavior is consistent with progressively relaxed thresholds.

For *person*, risks increase moderately with α as coverage expands (DINOv2: coverage $0.25 \rightarrow 0.60$ with risk $0.0803 \rightarrow 0.1400$; Mask2Former: coverage $0.30 \rightarrow 0.68$ with risk $0.0722 \rightarrow 0.1317$), suggesting a clear coverage-risk trade-off. The results show that epistemic and aleatoric thresholds vary more consistently with α, as no model exhibits the near-constant threshold behavior.

B Model-specific and Shared Experimental Configuration

Table 9 summarizes the model-specific and shared settings used throughout the evaluation. It highlights that the three pipelines follow a common three-stage training and calibration procedure, including weighted cross-entropy pretraining, focal-loss refinement, evidential fine-tuning, and class-conditional conformal calibration. At the same time, the pipelines differ in backbone architecture, segmentation head, optimization hyperparameters, and a small number of post-processing choices. This organization is important for interpreting the results, as it shows that the comparison is conducted under a largely uniform uncertainty-and-risk-control protocol, with only architecture-dependent adaptations introduced where required for stable training or calibration.

Table 9 Model configurations, training setup, and conformal calibration parameters with evidential fine-tuning. Shaded rows indicate parameters shared across all models. Focal-loss γ (Stage 2) is distinct from evidential γ_{EDL} (variance weight).

	Parameter	DINOv2	Mask2Former	SegFormer
Architecture	Backbone	ViT-Base `vit_base_patch14_dinov2` (pretrained)	SwinV2-L (pretrained)	MiT-B5 (pretrained)
	Head	1×1 Conv ($768 \to C$)	Mask2Former decoder + seg. head	MLP decoder head
	Output (Stage 3)	logits → evidence		
Training	Stage 1: epochs / bs / sched.	50 / 2 / cosine		
	Optimizer / WD / LR	AdamW; 10^{-4}; 3×10^{-5}	AdamW; 0.05; 1×10^{-4}	AdamW; 0.01; 6×10^{-5}
	Loss (Stage 1)	weighted CE		
	Stage 2: epochs / loss	40 / focal γ_{focal}=2.0	40 / focal γ_{focal}=2.0	40 / focal γ_{focal}=1.5
	Stage 3: epochs / loss	30 / CE+DEL		
	DEL params	λ_{DEL}=0.5; T_{anneal}=10; $\beta(e)=\min(1, e/10)$		
Calibration	Cal. split	20% held-out val, stratified by image	20% held-out val	20% held-out val
	Target α	$\{0.01, 0.02, 0.05, 0.10\}$		
	EDL weights	γ_{EDL}=1.0; $\beta(e)=\min(1, e/10)$		
	Temp. scaling	scalar T; init 1.0; bounds $[0.5, 5]$; LBFGS 50 iters	scalar T; init 1.0; bounds $[0.5, 6]$; LBFGS 50 iters	scalar T; init 1.0; bounds $[0.5, 4]$; LBFGS 50 iters
	Conformal score	pixel uncert. score $s(u)$	$s(u)$; post-upsampling	$s(u)$; stable for transformers
	Thresholds	class-cond. $\tau_{\hat{y}(u)}(\alpha)$; global fallback	class-cond. $\tau_{\hat{y}(u)}(\alpha)$; global fallback	class-cond. $\tau_{\hat{y}(u)}(\alpha)$; finite-sample corr.
	Min. support	$n_k \geq 2000$ px; else global τ	$n_k \geq 5000$ px; else global τ	$n_k \geq 2000$ px; else global τ
	Smoothing	unc.-mask opening 3×3; $A_{\min}$=25 px	–	median filter 3×3 on max-prob map
	Pred. cap	–	–	$K_{\max}$=3 labels/px; else "uncertain"
	Conf. bins (opt.)	–	3 bins by max-prob: $[0, 0.5)$, $[0.5, 0.8)$, $[0.8, 1]$; sep. $\tau_{k,b}$	–