Reliable Real-Time Edge AI via Conformal Model Selection

Anders E. Kalør and Tomoaki Ohtsuki
Department of Information and Computer Science, Keio University, Japan
E-mail: {aek, ohtsuki}@keio.jp

***Abstract*—Edge artificial intelligence (AI) is expected to be a central part of 6G, where servers located at the edge of the network will support devices in performing inference using machine learning (ML) models. However, providing latency and accuracy guarantees needed by many 6G applications, such as automated driving and robotics, is challenging due to the black-box nature of ML models, the complexity of the tasks, and the random wireless channel. This paper proposes a novel framework leveraging conformal risk control to meet requirements on the expected loss under a strict deadline. To adapt to fluctuating channel conditions, our framework utilizes an ensemble of black-box encoder/decoder models and inference models of varying accuracy and complexity, and selects the model expected to yield the most informative prediction under the given requirements. We demonstrate the proposed framework on a deadline-constrained image classification task under a strict missed detection requirement. The results suggest that the proposed framework provides the required performance guarantees, making it a promising step toward achieving reliable real-time edge AI services in 6G.**

I. Introduction

Driven by the success of artificial intelligence (AI), edge AI is expected to be a central component of 6G, where servers located at the edge of the network will support devices in performing inference and making decisions using machine learning (ML) [1]. For instance, edge servers may assist vehicles in performing image object detection in automated driving, or execute reinforcement learning models to control industrial robots. Such edge AI applications often operate under strict performance and time constraints, requiring inference results to be both accurate and delivered before a specific deadline with high probability.

Meeting these requirements involves trade-offs between the quality of transmitted data representations (affecting accuracy and uplink time), the computational complexity of edge ML models (affecting accuracy and processing time), and the size of the resulting predictions (affecting downlink transmission). Several techniques have been proposed to optimize these trade-offs, relying on, e.g., split inference [2]–[4] and over-the-air computing [5]. However, the analysis of these techniques is typically based on oversimplified data models, which rarely reflect practical settings, and rely on *white-box* ML models and complex feature extraction at the device. On the other hand, many practical ML models are either inherently black-box or too complex for white-box analysis. Furthermore, resource-constrained devices may not be able to compute complex features, relying instead on simple processing tools, such as image compression with various quality settings. Compared to a fixed feature vector, compression algorithms typically generate outputs of variable lengths, which influence the transmission delay.

In this paper, we present a generalized framework for black-box model selection that provides statistical guarantees on the resulting end-to-end loss and latency, accounting for challenges such as random message lengths. Given a loss function and a deadline, our framework jointly selects the transmission quality, by choosing from an ensemble of available black-box encoder/decoder pairs with varying complexities and execution times, and the inference model, by choosing from an ensemble of black-box ML models hosted on the edge server (see Fig. 1).

To provide statistically sound end-to-end guarantees for such black-box systems, our framework introduces a novel approach that combines conformal risk control to meet the loss requirement with non-parametric statistics to bound the delay violation probability while considering the random message lengths and channel conditions. The key idea behind conformal risk control [6], [7] is to output a *prediction set* rather than a single point estimate. Through careful calibration, conformal risk control aims to construct the smallest prediction set such that the expected loss is bounded by a predefined constant, thereby providing reliable predictions and quantifiable uncertainty estimates. Conformal risk control has previously been applied to various ML-based applications in wireless communication [8]–[10]. In addition to the reliability guarantees provided by conformal risk control, our framework relies on order statistics derived from an unlabeled dataset to obtain a distribution-free probabilistic bound on the joint sizes of the encoded uplink message and the prediction set resulting from applying conformal prediction to a given model combination. This is then combined with the analytical channel model to derive an end-to-end delay guarantee.

The remainder of the paper is organized as follows. Section II introduces the system model and formalizes the problem statement. The proposed schemes are presented in Section III and the numerical results are presented in Section IV. Finally, the paper is concluded in Section V.

II. System Model and Problem Statement

A. Scenario and Inference Model

We consider a deadline-constrained sensor connected wirelessly to an edge server (Fig. 1), operating in time frames of duration T. In each frame $t = 1, 2, \ldots$, the sensor observes an input $X_t \in \mathcal{X}$, which it offloads to the edge server for inference. Associated with the input X_t is an *unobservable* ground-truth label $Y_t \in \mathcal{Y}$, where $\mathcal{Y}$ is a discrete set of labels.

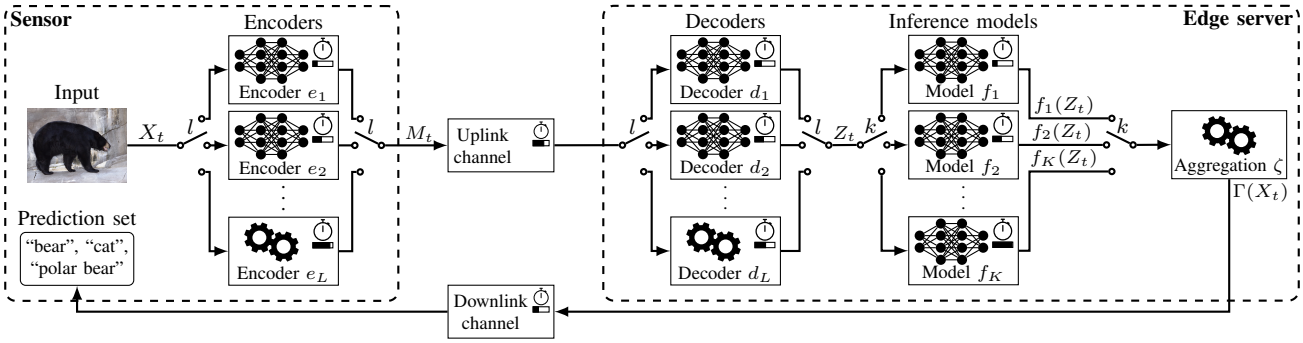

Fig. 1. The considered scenario. The sensor encodes its input X_t using a selected encoder as $M_t = e_l(X_t)$ and transmits it to the edge server. The server decodes the message as $Z_t = d_l(M_t)$ using the corresponding decoder and performs inference using a selected inference model, f_k. The model outputs $f_k(Z_t)$ are then aggregated into a prediction set $\Gamma(X_t) = f_k(Z_t)$, which must be transmitted back to the sensor before the deadline.

We assume that (X_t, Y_t) are drawn independently from an unknown joint distribution P_{XY}.

The sensor uses one of L encoders pairs $e_l : \mathcal{X} \to \{0,1\}^*$ to map input X_t to a message $M_t = e_l(X_t)$ of (variable) length $D_{\mathrm{ul},l}(X_t) = |M_t|$ bits, which it transmits to the edge server. The edge server decodes the message using the corresponding decoder $d_l : \{0,1\}^* \to \mathcal{Z}$ into an intermediate representation $Z_t = d_l(M_t) \in \mathcal{Z}$. Z_t serves as a common input representation for all subsequent edge inference models (e.g., a reconstructed image or a feature vector). The l-th encoder/decoder pair has a deterministic total computation time $\tau_{\mathrm{ul},l}$, comprising both encoding and decoding but excluding transmission delay, and offers a trade-off between message size, computing time, and representation quality.

The edge server performs inference on the received Z_t. Similar to the encoder/decoder configuration, we assume that the edge server performs inference using one of K pre-trained, black-box inference models $\{f_k\}_{k=1}^K$. Each inference model $f_k : \mathcal{Z} \to \mathbb{R}^{|\mathcal{Y}|}$, $k = 1, \ldots, K$, takes the representation $Z_t \in \mathcal{Z}$ as input (regardless of the encoder/decoder pair used for transmission) and outputs a confidence score $[f_k(Z_t)]_y \in \mathbb{R}$ of each label $y \in \mathcal{Y}$, e.g., using the softmax activation function. The k-th model has a fixed computation time τ_{f_k}. Typically, a model with a longer computation time is expected to produce better predictions, but this may not always be the case. In practice, the models may be implemented as different *scales* of the same architecture, each with a different number of layers, neurons, etc. [11]. However, we emphasize that our proposed framework is agnostic to the specific architecture of the underlying ML models, treating them effectively as black-box models.

For the selected edge model f_k, the edge server constructs a *prediction set* $\Gamma(X_t) \subseteq \mathcal{Y}$ by applying an aggregation function ζ to the confidence scores produced by the selected model:

$$\Gamma(X_t) = \zeta\left(f_k(Z_t)\right).$$

Note that the prediction set $\Gamma(X_t)$ contains a set of labels rather than a single point estimate. For instance, ζ could select a certain number of labels with the largest confidence scores, or all labels with a score greater than some threshold. In general, ζ controls the trade-off between *coverage* (i.e., the probability that it contains a the ground-truth) and *informativeness* (i.e., the size of the prediction set). Note that $\Gamma(X_t)$ is a random subset of $\mathcal{Y}$ that depends on the random input X_t through Z_t.

We assume that each predicted label $y \in \Gamma(X_t)$ occupies D_{lbl} bits, so that the size of the prediction set is

$$D_{\mathrm{dl},l,k}(X_t) = |\Gamma(X_t)| D_{\mathrm{lbl}} \quad \text{(bits)}.$$

Since each label may include metadata such as bounding box coordinates, depth estimates, textual descriptions, etc., D_{lbl} can potentially span from a few bits to several hundred bytes depending on the application.

B. Communication Model

We consider a Rayleigh block-fading channel, in which the channel gain remains constant throughout the transmission and changes independently between transmissions. The communication rate in the uplink is then given by

$$R_{\mathrm{ul},t} = B \log_2\left(1 + |h_{\mathrm{ul},t}|^2 \mathsf{SNR}\right) \quad \text{(bits/s)}, \tag{1}$$

where B is the bandwidth in Hz, $h_{\mathrm{ul},t} \sim \mathcal{CN}(0,1)$ is the instantaneous uplink channel gain in frame t, and SNR is the average signal-to-noise ratio (SNR), which is known to both the sensor and the edge server. We assume that the instantaneous channel gain is revealed to the sensor *after* the encoder/decoder pair has been selected (e.g., estimated using pilot signals), so that only the statistics of $R_{\mathrm{ul},t}$ can be used to select the encoder/decoder. The total duration of the observation transmission can then be computed as

$$T_{\mathrm{ul},t} = \tau_{\mathrm{ul},l} + \frac{D_{\mathrm{ul},l}(X_t)}{R_{\mathrm{ul},t}}, \tag{2}$$

where for simplicity we neglect the impact of protocol overhead.

Similar to the uplink, the rate in the downlink is

$$R_{\mathrm{dl},t} = B \log_2\left(1 + |h_{\mathrm{dl},t}|^2 \mathsf{SNR}\right) \quad \text{(bits/s)}, \tag{3}$$

so that the edge inference and downlink transmission time can be computed as

$$T_{\mathrm{dl},t} = \tau_{f_k} + \frac{D_{\mathrm{dl},l,k}(X_t)}{R_{\mathrm{dl},t}}. \tag{4}$$

As in the uplink, we assume that the instantaneous rate cannot be used to select the edge model f_k. However, we will also

consider a variant of the problem, detailed in Section III-E, where knowledge of the supported rate can be used to truncate the prediction set $\Gamma(X_t)$ to fit within the frame.

C. Problem Statement

The prediction quality is characterized by a loss function $\ell(\Gamma(X_t), Y_t)$. For technical reasons, we assume that the loss can never increase by enlarging $\Gamma(X_t)$, and that it is upper bounded by some constant γ. Note that these conditions are satisfied for many common loss functions, such as the 0-1 loss and the missed detection probability. Using this, we define the *risk-adjusted loss* ℓ', which assigns a loss of ℓ whenever the deadline is met and is otherwise equal to γ, i.e.,

$$\ell'\left(\Gamma(X_t), Y_t\right) = \begin{cases} \ell(\Gamma(X_t), Y_t), & \text{if } T_{\text{tot},t} \leq T, \\ \gamma, & \text{otherwise}, \end{cases} \tag{5}$$

where $T_{\text{tot},t} = T_{\text{ul},t} + T_{\text{dl},t}$ is the round-trip inference duration.

Our goal is to jointly select the encoder/decoder, the edge inference model, and the aggregation function ζ that achieve the most informative (i.e., smallest) prediction set $\Gamma(X_t)$ while having an expected risk-adjusted loss of at most α. Specifically, we aim to solve

$$\begin{aligned} \text{minimize} \quad & \mathbb{E}\left[|\Gamma(X_t)| \mid T_{\text{tot},t} \leq T\right], && \text{(6a)} \\ \text{s.t.} \quad & \mathbb{E}\left[\ell'\left(\Gamma(X_t), Y_t\right)\right] \leq \alpha, && \text{(6b)} \end{aligned}$$

where the expectations are over $(X_t, Y_t) \sim P_{XY}$ and $h_{\text{ul},t}, h_{\text{dl},t} \sim \mathcal{CN}(0,1)$.

Solving Problem (6) optimally is generally challenging since P_{XY} is unknown. Instead, we assume access to labeled and unlabeled *calibration* datasets. The labeled dataset is denoted by $\mathcal{D}$ and contains $N_{\mathcal{D}}$ samples drawn independently and identically distributed (i.i.d.) from P_{XY}, i.e.,

$$\mathcal{D} = \{(X_n^{(\mathcal{D})}, Y_n^{(\mathcal{D})})\}_{n=1}^{N_{\mathcal{D}}}, \quad (X_n^{(\mathcal{D})}, Y_n^{(\mathcal{D})}) \overset{\text{i.i.d.}}{\sim} P_{XY}. \tag{7}$$

Similarly, the unlabeled dataset, denoted by $\mathcal{U}$, contains $N_{\mathcal{U}}$ input samples drawn i.i.d. from the marginal input distribution, P_X, of P_{XY} and independently from $\mathcal{D}$:

$$\mathcal{U} = \{X_n^{(\mathcal{U})}\}_{n=1}^{N_{\mathcal{U}}}, \quad X_n^{(\mathcal{U})} \overset{\text{i.i.d.}}{\sim} P_X. \tag{8}$$

Utilizing these datasets, we seek a model selection procedure that satisfy the loss requirement on *unseen samples* drawn from P_{XY}, while minimizing the prediction set size.

III. End-to-End Conformal Model Selection

This section presents a general framework for jointly optimizing the choice of encoder/decoder and inference models to solve Problem (6). The framework operates in four steps:

1) We first separate Constraint (6b) into distinct loss and frame deadline requirements;
2) We employ conformal risk control to define the aggregation function ζ, calibrating it on a specific dataset for each composite encoder/decoder and inference model combination to satisfy the new loss requirement;
3) Given the calibrated aggregation function ζ, we compute the probability that each model meets the frame deadline and discard the model combinations that do not satisfy the requirement;
4) Finally, we estimate the expected prediction set size of each remaining model combination and select the one that maximizes the expected informativeness.

A. Constraint Separation

We decompose Constraint (6b) into separate loss and frame deadline requirements that can be addressed individually, while ensuring that satisfying the separate requirements is a sufficient condition for satisfying the original constraint. To simplify the notation, let $\ell = \ell\left(\Gamma(X_t), Y_t\right)$, $\ell' = \ell'\left(\Gamma(X_t), Y_t\right)$, $T_{\leq T} = T_{\text{tot},t} \leq T$, and $T_{>T} = T_{\text{tot},t} > T$. The risk-adjusted loss in Eq. (5) can then be bounded as

$$\begin{aligned} \mathbb{E}\left[\ell'\right] &= \mathbb{E}\left[\ell \mid T_{\leq T}\right] \Pr(T_{\leq T}) + \gamma \Pr(T_{>T}) \\ &= \mathbb{E}\left[\ell\right] + \left(\gamma - \mathbb{E}\left[\ell \mid T_{>T}\right]\right) \Pr(T_{>T}) \\ &\leq \mathbb{E}\left[\ell\right] + \gamma \Pr(T_{>T}), \end{aligned}$$

where the equality is obtained by substituting $\mathbb{E}\left[\ell | T_{\leq T}\right] \Pr(T_{\leq T}) = \mathbb{E}\left[\ell\right] - \mathbb{E}\left[\ell | T_{>T}\right] \Pr(T_{>T})$ using the law of total expectation and rearranging, and the inequality follows since $(\gamma - \mathbb{E}\left[\ell | T_{>T}\right])$ is upper bounded by γ.

To ensure that Constraint (6b) is satisfied, it is sufficient to require $\mathbb{E}\left[\ell\right] + \gamma \Pr(T_{>T}) \leq \alpha$. This can be achieved by bounding the first and second terms by $\delta\alpha$ and $(1-\delta)\alpha$, respectively, for any $0 \leq \delta \leq 1$. This gives us the separate loss and frame deadline requirements

$$\mathbb{E}\left[\ell\left(\Gamma(X_t), Y_t\right)\right] \leq \delta\alpha, \tag{9}$$

$$\Pr(T_{\text{tot},t} > T) \leq \frac{(1-\delta)\alpha}{\gamma}, \tag{10}$$

which are sufficient conditions for Constraint (6b). δ controls the trade-off between the expected loss $\mathbb{E}\left[\ell\left(\Gamma(X_t), Y_t\right)\right]$ and the scaled deadline violation probability $\gamma \Pr(T_{\text{tot},t} > T)$.

B. Bounding the Expected Loss using Conformal Risk Control

In this section, we use conformal risk control [7] to design the aggregation function ζ such that the requirement in Eq. (9) is satisfied. Conformal risk control belongs to the conformal prediction framework [6], which provides model-agnostic, distribution-free statistical guarantees of ML model predictions. The central idea is to use a threshold-based aggregation function

$$\zeta\left(f_k(Z_t)\right) = \left\{y \in \mathcal{Y} : \left[f_k(Z_t)\right]_y \geq 1 - \lambda\right\}, \tag{11}$$

and then carefully calibrate the confidence score threshold λ for each combination of an encoder/decoder and an edge inference model using the calibration dataset $\mathcal{D}$. To this end, we define the composite model comprising encoder/decoder pair (e_l, d_l) and edge inference model f_k as

$$g_{l,k}(X) = f_k\left(d_l\left(e_l(X)\right)\right), \tag{12}$$

for all $l = 1, 2, \ldots, L$ and $k = 1, 2, \ldots, K$, and denote the confidence score threshold for composite model $g_{l,k}$ as $\lambda_{l,k}$. The following lemma shows how to select $\lambda_{l,k}$ to satisfy $\mathbb{E}[\ell(\Gamma_{\lambda_{l,k},l,k}(X), Y)] \leq \varepsilon$.

Lemma 1 (Conformal risk control [6], [7]): Let $\mathcal{D}$ be defined as in Eq. (7), and let $\Gamma_{\lambda,l,k}(x)$ denote the prediction set constructed on input x using the aggregation function in (11) for a composite model $g_{l,k}$ with the threshold λ. Suppose the loss function ℓ satisfies

$$\ell(\Gamma_{\lambda_2,l,k}(x),y) \leq \ell(\Gamma_{\lambda_1,l,k}(x),y) \leq \gamma$$

for all (x,y) and $\lambda_1 \leq \lambda_2$, and for some finite γ. The threshold

$$\lambda_{l,k} = \inf\left\{\lambda : \frac{1}{N_{\mathcal{D}}}\sum_{n=1}^{N_{\mathcal{D}}} \ell(\Gamma_{\lambda,l,k}(X_n^{(\mathcal{D})}), Y_n^{(\mathcal{D})}) \leq \varepsilon - \frac{\gamma-\varepsilon}{N_{\mathcal{D}}}\right\}, \tag{13}$$

then satisfies

$$\mathbb{E}_{(X,Y)\sim P_{XY}}\left[\ell\left(\Gamma_{\lambda_{l,k},l,k}(X),Y\right)\right] \leq \varepsilon.$$

Selecting the thresholds based on Eq. (13) with $\varepsilon = \delta\alpha$ thus ensures that Eq. (9) is met for all model combinations. Note that this approach of calibrating each model combination is necessary since the intermediate representations can vary significantly between different encoder/decoder pairs. Consequently, a calibration that works well for one encoder/decoder pair might be ineffective for another, necessitating individual calibration of each $g_{l,k}$. However, because the number of model combinations is typically small and calibration is performed offline, this approach remains practical.

C. Bounding the Deadline Violation Probability

The technique presented in Section III-B allows us to guarantee that Constraint (9) is satisfied. In this section, we consider the problem of satisfying Constraint (10). Besides the model execution times, the round-trip delay of a given composite model depends on the random length of the message produced by the encoder, the size of the prediction set given the threshold obtained in Section III-B, and on the instantaneous uplink and downlink channels.

Bounding Eq. (10) is challenging due to the unknown, potentially dependent message length $D_{\mathrm{ul},l}(X_t)$ and prediction set size $|\Gamma(X_t)|$. For instance, a long message may be more likely to produce a small prediction set, or vice versa. On the other hand, the channel rates can be characterized analytically. Borrowing ideas from conformal risk control, in this paper we propose a technique to bound the end-to-end delay violation probability that uses the unlabeled dataset $\mathcal{U}$ to statistically bound the random message length and the prediction set size, and then combines these bounds with the analytical characterization of the channel rates. The resulting bound is presented in Proposition 1, and can be computed after the thresholds $\lambda_{l,k}$ have been determined as outlined in Section III-B.

Proposition 1 (Delay violation bound): Consider a composite model $g_{l,k}$ as defined in Eq. (12). Let $\sigma_{\mathrm{ul},l}$ and $\sigma_{\mathrm{dl},l,k}$ be index permutations that order the unlabeled dataset $\mathcal{U}$ based on the size of the uplink and downlink data, respectively, i.e.,

$$D_{\mathrm{ul},l}(X^{(\mathcal{U})}_{\sigma_{\mathrm{ul},l}(1)}) \leq \ldots \leq D_{\mathrm{ul},l}(X^{(\mathcal{U})}_{\sigma_{\mathrm{ul},l}(N_{\mathcal{U}})}),$$
$$D_{\mathrm{dl},l,k}(X^{(\mathcal{U})}_{\sigma_{\mathrm{dl},l,k}(1)}) \leq \ldots \leq D_{\mathrm{dl},l,k}(X^{(\mathcal{U})}_{\sigma_{\mathrm{dl},l,k}(N_{\mathcal{U}})}).$$

The delay violation probability is then upper bounded as

$$\Pr\left(T_{\mathrm{tot},t} > T | g_{l,k}\right) \leq \min_{n,m\in\{1,\ldots,N_{\mathcal{U}}\}} 1 - e^{\bar{\beta}_{\mathrm{cal}}(n,m)}\left(\frac{n+m}{N_{\mathcal{U}}+1} - 1\right),$$

where

$$\bar{\beta}_{\mathrm{cal}}(n,m) = \frac{1}{2\mathsf{SNR}}\left(1 - 2^{\frac{\bar{D}_{\mathrm{ul},l}(n)+\bar{D}_{\mathrm{dl},l,k}(m)}{B\left(T-\tau_{\mathrm{ul},l}-\tau_{f_k}\right)}}\right), \tag{14}$$

and

$$\bar{D}_{\mathrm{ul},l}(n) = D_{\mathrm{ul},l}(X^{(\mathcal{U})}_{\sigma_{\mathrm{ul},l}(n)}), \tag{15}$$

$$\bar{D}_{\mathrm{dl},l,k}(m) = D_{\mathrm{dl},l,k}(X^{(\mathcal{U})}_{\sigma_{\mathrm{dl},l,k}(m)}), \tag{16}$$

are the n-th and m-th order statistics of $\{D_{\mathrm{ul},l}(X_n^{(\mathcal{U})})\}_{n=1}^{N_{\mathcal{U}}}$ and $\{D_{\mathrm{dl},l,k}(X_n^{(\mathcal{U})})\}_{n=1}^{N_{\mathcal{U}}}$, respectively.

Proof: See Appendix A. ∎

Computing the delay violation bound in Proposition 1 of each composite model $g_{l,k}$ with the thresholds determined in Section III-B enables us to discard the models that do not satisfy Constraint (10). The remaining models are guaranteed to satisfy the original Constraint (6b), and our final task is to select the best among those, which we consider next.

D. Informativeness Maximization and Model Selection

Once we have identified the set of models that satisfy Constraint (6b), we select the model for execution that maximizes the expected informativeness as specified by the objective in Eq. (6a). Since the conditional expectation in (6a) is hard to compute, we heuristically select the valid model $g_{l,k}$ (i.e., satisfying the bound from Proposition 1) with the minimum *unconditional* average prediction set size $\bar{\Gamma}_{l,k} = \frac{1}{N_{\mathcal{U}}}\sum_{n=1}^{N_{\mathcal{U}}} |\Gamma_{\lambda_{l,k},l,k}(X_n^{(\mathcal{U})})|$ estimated on $\mathcal{U}$. The complete procedure is summarized in the algorithm below.

function MODELSELECTION
  Initialize $g^* \leftarrow$ NULL; $\lambda^* \leftarrow 0$; $\bar{\Gamma}^* \leftarrow \infty$; $\bar{P}^* \leftarrow \infty$.
  for $l = 1, 2, \ldots, L$ **do**
    for $k = 1, 2, \ldots, K$ **do**
      Define $\Gamma_{\lambda,l,k}(X) = \{y \in \mathcal{Y} : [g_{l,k}(X)]_y \geq 1 - \lambda\}$.
      Compute $\lambda_{l,k}$ for $\Gamma_{\lambda,l,k}$ using Lemma 1 with $\mathcal{D}$ and $\varepsilon = \delta\alpha$.
      Compute $\bar{P}_{l,k}$ as the bound in Prop. 1 using $\mathcal{U}$ and $\lambda_{l,k}$.
      $\bar{\Gamma}_{l,k} \leftarrow \frac{1}{N_{\mathcal{U}}}\sum_{n=1}^{N_{\mathcal{U}}} |\Gamma_{\lambda_{l,k},l,k}(X_n^{(\mathcal{U})})|$.
      if ($\bar{P}_{l,k} \leq (1-\delta)\alpha/\gamma$ *and* $\bar{\Gamma}_{l,k} < \bar{\Gamma}^*$)
        or ($\bar{P}_{l,k} \geq (1-\delta)\alpha/\gamma$ *and* $\bar{P}_{l,k} < \bar{P}^*$) **then**
        $g^* \leftarrow g_{l,k}$; $\lambda^* \leftarrow \lambda_{l,k}$; $\bar{\Gamma}^* \leftarrow \bar{\Gamma}_{l,k}$; $\bar{P}^* \leftarrow \bar{P}_{l,k}$.
  return (g^*, λ^*).

Note that, if none of the models satisfy Constraint (6b), we pick the model with the smallest delay violation probability. Furthermore, while the computational complexity might be significant when $N_{\mathcal{D}}$ and $N_{\mathcal{U}}$ are large, the procedure is intended to be computed offline as it only relies on the average SNR.

E. Channel-Aware Prediction Set Truncation

Until now, we have assumed that the prediction set is selected only based on thresholding. However, if the instantaneous downlink channel rate $R_{\mathrm{dl},t}$ is known at the server, the prediction set can be truncated to guarantee that it can

TABLE I
MODELS USED IN THE NUMERICAL RESULTS

Encoder/decoder, (e_l, d_l)	Quality setting	Execution time, $\tau_{\mathrm{ul},l}$
WebP-0, (e_1, d_1)	0	10.0 ms
WebP-20, (e_2, d_2)	20	12.5 ms
WebP-50, (e_3, d_3)	50	15.0 ms
WebP-80, (e_4, d_4)	80	17.5 ms
Classifier model, f_k	**Num. parameters**	**Execution time, $\tau_{\mathrm{dl},k}$**
EfficientNetV2-S, f_1	22M	24.0 ms
EfficientNetV2-M, f_2	54M	57.0 ms
EfficientNetV2-L, f_3	120M	98.0 ms

be delivered before the deadline. In such case, the edge server model may truncate the prediction set generated by the selected model $g_{l,k}$ to contain at most

$$\tilde{\Gamma}_t = \max\left(1, \left\lfloor \frac{R_{\mathrm{dl},t}(T-\tau_{\mathrm{ul},l}-\tau_{f_k}-T_{\mathrm{ul},t})}{D_{\mathrm{lbl}}} \right\rfloor\right),$$

so that the resulting prediction set is given as

$$\Gamma(X_t) = \left\{y \in \mathcal{Y} : [g_{l,k}(X_t)]_y \geq 1-\lambda_{l,k}, y \in \mathrm{top}_{\tilde{\Gamma}_t}(g_{l,k}(X_t))\right\},$$

where $\mathrm{top}_{\tilde{\Gamma}_t}(g_{l,k}(X_t))$ are the $\tilde{\Gamma}_t$ labels with the largest confidence scores. While the impact of truncation on the average set size (6a) varies, it never violates the risk constraint (6b), as it can only prevent deadline misses (which incur maximum loss γ).

IV. Numerical Results

Scenario: We demonstrate the proposed framework on an image classification task using the EfficientNetV2 [11] classifier models on the edge server pre-trained on the ImageNet 2012 dataset. The encoders/decoders are implemented using the WebP [12] image compression algorithm with various quality settings (see Table I). We evaluate the system on the ImageNet validation dataset comprising 50000 images of 1000 classes [13]. The dataset is randomly split into three disjoint sets for calibration ($N_{\mathcal{D}} = 10000$ labeled and $N_{\mathcal{U}} = 10000$ unlabeled) and evaluation (30000 labeled). We consider the 0-1 missed detection loss $\ell(\Gamma(X), Y) = \mathbb{1}[Y \notin \Gamma(X)]$, i.e., $\gamma = 1$, with requirement $\alpha = 0.02$ and deadline $T = 150$ ms. Each predicted label $y \in \mathcal{Y}$ is assumed to occupy $D_{\mathrm{lbl}} = 64$ bits, and the bandwidth is 30 MHz. Throughout, we set $\delta = 1/2$.

Baselines: We compare our proposed scheme to a small and a large fixed model policy. The small baseline model comprises the WebP-0 encoder/decoder and the EfficientNetV2-S classifier, i.e., $g_{1,1}$, while the large baseline model is defined by WebP-80 and EfficientNetV2-L, i.e., $g_{4,3}$. For each model, we consider top-20 and calibrated threshold-based aggregation presented in Section III-B.

Results: Figure 2 shows that our proposed schemes (solid blue and green) almost always meet the loss requirement $\alpha = 0.02$, except for very low SNR, where none of the model combinations satisfy the requirement. Consequently, the framework selects the model predicted to have the lowest delay violation probability, which still results in an average loss exceeding α. Both fixed models with top-20 aggregation (solid orange and purple) fail to meet the requirements. The

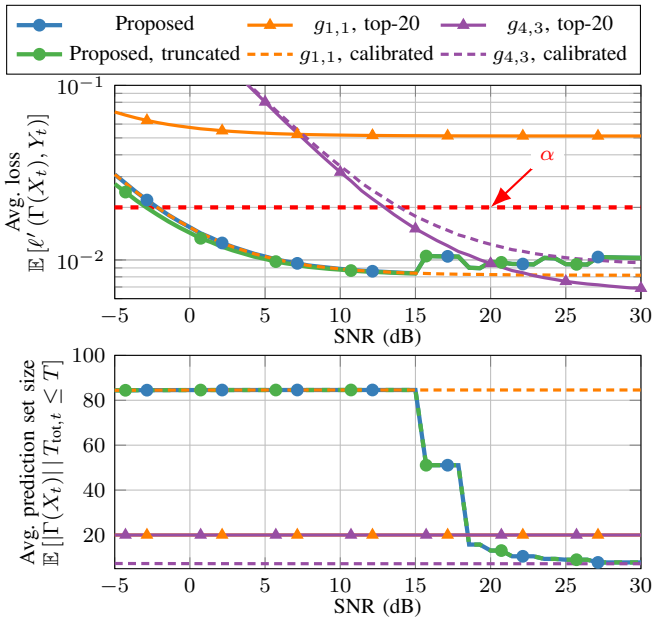

Fig. 2. Average loss (top) and prediction set size (bottom) vs. SNR.

small fixed calibrated model (dashed orange) meets the loss, but with a much larger average prediction set size at high SNR. The large fixed calibrated model (dashed purple) provides a smaller prediction set, but only satisfies the loss requirement at high SNR due to longer transmission and execution times. On the other hand, the prediction set size of the proposed models decrease steadily as the SNR increases while satisfying the loss requirement as desired. Prediction set truncation (dashed blue line) provides only a small benefit as uplink transmission delay dominates the total execution time.

V. Conclusion

In this paper, we present a framework for reliable real-time edge AI under strict loss and deadline requirements. We assume that the sensor and edge server have access to an ensemble of black-box encoder/decoder and inference models with various complexities and execution times. Using ideas from conformal risk control, we propose a model selection scheme that aims to maximize the informativeness of the predictions under bounded loss and deadline violation probability. Through numerical results of an image classification scenario, we demonstrate that the proposed framework meets the requirements while minimizing the average size of the prediction sets. This suggests that the proposed framework is a promising direction toward achieving reliable and timely edge AI services in 6G.

Appendix
Proof of Proposition 1

Let $D_{\mathrm{ul},t} = D_{\mathrm{ul},l}(X)$ and $D_{\mathrm{dl},t} = D_{\mathrm{dl},l,k}(X)$ where $X \sim P_X$. Then, for any $D'_{\mathrm{ul},t}, D'_{\mathrm{dl},t} \geq 0$,

$$\begin{aligned}
&\Pr(T_{\mathrm{tot},t} > T \,|\, g_{l,k}) \\
&\overset{(a)}{=} 1 - \int_0^\infty \int_0^\infty \Pr(T_{\mathrm{tot},t} \leq T \,|\, D_{\mathrm{ul},t} = \xi, D_{\mathrm{dl},t} = \psi, g_{l,k}) \\
&\qquad \times p(D_{\mathrm{ul},t} = \xi, D_{\mathrm{dl},t} = \psi \,|\, g_{l,k})\, \mathrm{d}\xi\, \mathrm{d}\psi
\end{aligned}$$

$$\overset{(b)}{\leq} 1-\int_0^{D'_{\mathrm{ul},t}}\int_0^{D'_{\mathrm{dl},t}} \Pr(T_{\mathrm{tot},t}\leq T|D_{\mathrm{ul},t}=\xi, D_{\mathrm{dl},t}=\psi, g_{l,k}) \times p(D_{\mathrm{ul},t}=\xi, D_{\mathrm{dl},t}=\psi \,|\, g_{l,k})\,\mathrm{d}\xi\,\mathrm{d}\psi$$
$$\overset{(c)}{\leq} 1-\Pr(T_{\mathrm{tot},t}\leq T \,|\, D_{\mathrm{ul},t}=D'_{\mathrm{ul},t}, D_{\mathrm{dl},t}=D'_{\mathrm{dl},t}, g_{l,k}) \times \int_0^{D'_{\mathrm{ul},t}}\int_0^{D'_{\mathrm{dl},t}} p(D_{\mathrm{ul},t}=\xi, D_{\mathrm{dl},t}=\psi|g_{l,k})\,\mathrm{d}\xi\,\mathrm{d}\psi$$
$$\overset{(d)}{=} 1-\Pr(T_{\mathrm{tot},t}\leq T \,|\, D_{\mathrm{ul},t}=D'_{\mathrm{ul},t}, D_{\mathrm{dl},t}=D'_{\mathrm{dl},t}, g_{l,k}) \times \Pr(D_{\mathrm{ul},t}\leq D'_{\mathrm{ul},t}, D_{\mathrm{dl},t}\leq D'_{\mathrm{dl},t} \,|\, g_{l,k}). \tag{17}$$

Here, (a) follows from the law of total probability, and (b) follows from the non-negativity of the CDF. Step (c) is due to the fact that $\Pr(T_{\mathrm{tot},t}\leq T \,|\, D_{\mathrm{ul},t}, D_{\mathrm{dl},t}, g_{l,k})$ is nonincreasing in $D_{\mathrm{ul},t}$ and $D_{\mathrm{dl},t}$, and (d) is obtained by noting that the integral evaluates to the joint CDF.

We proceed to establish a lower bound on $\Pr(T_{\mathrm{tot},t} \leq T \,|\, D_{\mathrm{ul},t}, D_{\mathrm{dl},t}, g_{l,k})$. For any $0\leq\phi\leq T$,

$$\begin{aligned}
&\Pr(T_{\mathrm{tot},t}\leq T \,|\, D_{\mathrm{ul},t}, D_{\mathrm{dl},t}, g_{l,k})\\
&= \Pr(T_{\mathrm{ul},t}+T_{\mathrm{dl},t}\leq T \,|\, D_{\mathrm{ul},t}, D_{\mathrm{dl},t}, g_{l,k})\\
&\geq \Pr(T_{\mathrm{ul},t}\leq T-\phi, T_{\mathrm{dl},t}\leq \phi \,|\, D_{\mathrm{ul},t}, D_{\mathrm{dl},t}, g_{l,k})\\
&= \Pr(T_{\mathrm{ul},t}\leq T-\phi|D_{\mathrm{ul},t}, g_{l,k})\Pr(T_{\mathrm{dl},t}\leq \phi|D_{\mathrm{dl},t}, g_{l,k}),
\end{aligned} \tag{18}$$

where the inequality follows from the fact that $T_{\mathrm{ul},t}$ and $T_{\mathrm{dl},t}$ are conditionally independent given $D_{\mathrm{ul},t}, D_{\mathrm{dl},t}, g_{l,k}$, and that $T_{\mathrm{ul},t}$ is independent of $D_{\mathrm{dl},t}$, while $T_{\mathrm{dl},t}$ is independent of $D_{\mathrm{ul},t}$. Expanding the terms first using (2) and (4), and then using (1) and (3) while using that $|h_{\mathrm{ul},t}|^2$ and $|h_{\mathrm{dl},t}|^2$ are exponentially distributed yields

$$\begin{aligned}
&\Pr(T_{\mathrm{ul},t}\leq T-\phi \,|\, D_{\mathrm{ul},t}, g_{l,k})\Pr(T_{\mathrm{dl},t}\leq \phi \,|\, D_{\mathrm{dl},t}, g_{l,k})\\
&= \Pr\left(R_{\mathrm{ul},t}\geq \tfrac{D_{\mathrm{ul},t}}{T-\phi-\tau_{\mathrm{ul},l}}\middle| D_{\mathrm{ul},t}\right)\Pr\left(R_{\mathrm{dl},t}\geq \tfrac{D_{\mathrm{dl},t}}{\phi-\tau_{f_k}}\middle| D_{\mathrm{dl},t}\right)\\
&= \exp\left(\frac{1-2^{\frac{D_{\mathrm{ul},t}}{B(T-\phi-\tau_{\mathrm{ul},l})}}}{\mathsf{SNR}}\right)\exp\left(\frac{1-2^{\frac{D_{\mathrm{dl},t}}{B(\phi-\tau_{f_k})}}}{\mathsf{SNR}}\right)\\
&= \exp\left(\frac{2-2^{\frac{D_{\mathrm{ul},t}}{B(T-\phi-\tau_{\mathrm{ul},l})}}-2^{\frac{D_{\mathrm{dl},t}}{B(\phi-\tau_{f_k})}}}{\mathsf{SNR}}\right),
\end{aligned} \tag{19}$$

Setting $\phi=(D_{\mathrm{ul},t}\tau_{f_k}+D_{\mathrm{dl},t}(T-\tau_{\mathrm{ul},l}))/(D_{\mathrm{ul},t}+D_{\mathrm{dl},t})$, and substituting this into (19) gives

$$\begin{aligned}
&\Pr(T_{\mathrm{tot},t}\leq T \,|\, D_{\mathrm{ul},t}, D_{\mathrm{dl},t}, g_{l,k})\\
&\quad\geq \exp\left(\tfrac{1}{2\mathsf{SNR}}\left(1-2^{\frac{D_{\mathrm{ul},t}+D_{\mathrm{dl},t}}{B(T-\tau_{\mathrm{ul},l}-\tau_{f_k})}}\right)\right).
\end{aligned} \tag{20}$$

Next, we bound $\Pr(D_{\mathrm{ul},t}\leq D'_{\mathrm{ul},t}, D_{\mathrm{dl},t}\leq D'_{\mathrm{dl},t} \,|\, g_{l,k})$ in (17). This quantity does not have an analytical expression, as it depends on the black-box encoder/decoder models (e_l, d_l), the edge model f_k, and the unknown distribution P_X. Instead, we bound it using the unlabeled dataset $\mathcal{U}$. By the inclusion-exclusion principle,

$$\begin{aligned}
&\Pr(D_{\mathrm{ul},t}\leq D'_{\mathrm{ul},t}, D_{\mathrm{dl},t}\leq D'_{\mathrm{dl},t} \,|\, g_{l,k})\\
&\geq \Pr(D_{\mathrm{ul},t}\leq D'_{\mathrm{ul},t}|g_{l,k}) + Pr(D_{\mathrm{dl},t}\leq D'_{\mathrm{dl},t}|g_{l,k}) - 1.
\end{aligned} \tag{21}$$

Conditioned on the threshold $\lambda_{l,k}$ and the model choice $g_{l,k}$, the marginal data size samples $\{D_{\mathrm{ul},l}(X_n^{(\mathcal{U})})\}_{n=1}^{N_{\mathcal{U}}}$ and $\{D_{\mathrm{dl},l,k}(X_n^{(\mathcal{U})})\}_{n=1}^{N_{\mathcal{U}}}$ are each a collection of independent samples drawn from the marginal distributions $p(D_{\mathrm{ul},t} \,|\, g_{l,k})$ and $p(D_{\mathrm{dl},t} \,|\, g_{l,k})$, respectively. Thus, the data sizes $D_{\mathrm{ul},t}$ and $D_{\mathrm{dl},t}$ of a new sample drawn from P_X are equally likely to fall anywhere between the samples in the dataset, i.e.,

$$\Pr(D_{\mathrm{ul},t}\leq \bar{D}_{\mathrm{ul},l}(n) \,|\, g_{l,k}) = \frac{n}{N_{\mathcal{U}}+1},$$
$$\Pr(D_{\mathrm{dl},t}\leq \bar{D}_{\mathrm{dl},l,k}(m) \,|\, g_{l,k}) = \frac{m}{N_{\mathcal{U}}+1},$$

for any integers $n, m \in \{1,\ldots,N_{\mathcal{U}}\}$, where $\bar{D}_{\mathrm{ul},l}(n)$ and $\bar{D}_{\mathrm{dl},l,k}(m)$ are defined as in Eqs. (15) and (16) (see, e.g., [6, Appendix D]). Combining this result with Eq. (21) yields

$$\begin{aligned}
&\Pr(D_{\mathrm{ul},t}\leq \bar{D}_{\mathrm{ul},l}(n), D_{\mathrm{dl},t}\leq \bar{D}_{\mathrm{dl},l,k}(m) \,|\, g_{l,k})\\
&\geq \tfrac{n}{N_{\mathcal{U}}+1}+\tfrac{m}{N_{\mathcal{U}}+1}-1\\
&= \tfrac{n+m}{N_{\mathcal{U}}+1}-1.
\end{aligned} \tag{22}$$

for any $n, m \in \{1,\ldots,N_{\mathcal{U}}\}$.

The proof is completed by inserting (20) and (22) into (17) with $D'_{\mathrm{ul},t}=\bar{D}_{\mathrm{ul},l}(n)$ and $D'_{\mathrm{dl},t}=\bar{D}_{\mathrm{dl},l,k}(m)$, defining $\bar{\beta}_{\mathrm{cal}}(l,k,n,m)$ as in Eq. (14), and choosing n and m such that the bound is minimized.

ACKNOWLEDGMENT

This work was supported by the Japan Science and Technology Agency (JST) ASPIRE program (grant no. JPMJAP2326) and by the Mizuho Foundation for the Promotion of Sciences.

REFERENCES

[1] A. E. Kalør *et al.*, "Wireless 6G connectivity for massive number of devices and critical services," *Proc. IEEE*, 2024, early access.

[2] E. Li, L. Zeng, Z. Zhou, and X. Chen, "Edge AI: On-demand accelerating deep neural network inference via edge computing," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 447–457, 2019.

[3] Z. Wang, A. E. Kalør, Y. Zhou, P. Popovski, and K. Huang, "Ultra-low-latency edge inference for distributed sensing," *IEEE Trans. Wireless Commun.*, 2025, early access.

[4] Q. Zeng, J. Huang, Z. Wang, K. Huang, and K. K. Leung, "Ultra-low-latency edge intelligent sensing: A source-channel tradeoff and its application to coding rate adaptation," *arXiv:2503.04645*, 2025.

[5] Z. Liu, Q. Lan, A. E. Kalør, P. Popovski, and K. Huang, "Over-the-air multi-view pooling for distributed sensing," *IEEE Trans. Wireless Commun.*, vol. 23, no. 7, pp. 7652–7667, 2024.

[6] A. N. Angelopoulos and S. Bates, "Conformal prediction: A gentle introduction," *Found. Trends Mach. Learn*, vol. 16, no. 4, pp. 494–591, 2023.

[7] A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, and T. Schuster, "Conformal risk control," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.

[8] K. M. Cohen, S. Park, O. Simeone, P. Popovski, and S. Shamai, "Guaranteed dynamic scheduling of ultra-reliable low-latency traffic via conformal prediction," *IEEE Signal Process. Lett.*, vol. 30, pp. 473–477, 2023.

[9] K. M. Cohen, S. Park, O. Simeone, and S. Shamai Shitz, "Calibrating AI models for wireless communications via conformal prediction," *IEEE Trans. Mach. Learn. Commun. Netw.*, vol. 1, pp. 296–312, 2023.

[10] M. Zhu, M. Zecchin, S. Park, C. Guo, C. Feng, and O. Simeone, "Federated inference with reliable uncertainty quantification over wireless channels via conformal prediction," *IEEE Trans. Signal Process.*, vol. 72, pp. 1235–1250, 2024.

[11] M. Tan and Q. Le, "EfficientNetV2: Smaller models and faster training," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 10 096–10 106.

[12] J. Zern, P. Massimino, and J. Alakuijala, "WebP image format," Nov. 2024. [Online]. Available: https://www.rfc-editor.org/info/rfc9649

[13] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vision*, vol. 115, no. 3, pp. 211–252, 2015.