When to Trust Beam Prediction: Confidence-Aware Adaptive Beam Management with Conformal Guarantees

Johan Eliasson
CTO, Gazella

***Abstract*—Machine learning for millimeter-wave (mmWave) beam prediction reduces measurement overhead but provides no reliability guarantees—a misaligned beam can cause severe throughput loss. We propose a confidence-aware adaptive beam management framework with three contributions: (1) a cost-aware error analysis showing that beam prediction errors have highly non-uniform severity, with off-by-1 errors losing <1 dB but off-by-5+ errors losing >5 dB; (2) conformal prediction sets—including a novel beam-aware variant exploiting spatial beam structure—that provide mathematically guaranteed $\geq 90\%$ coverage of the true beam; and (3) an adaptive fallback protocol that uses ML-predicted beams when confident and reverts to exhaustive search when uncertain. We evaluate six methods spanning logistic regression, MLPs, residual networks, 1D-CNNs, and self-attention on a distance-dependent clustered channel at 28 GHz with 64 narrow beams. Over three random seeds, the adaptive system achieves near-100% effective accuracy while reducing average measurement overhead by 40–60% at moderate confidence thresholds ($K \leq 5$). The beam-aware conformal method produces tighter, spatially contiguous prediction sets compared to standard conformal calibration.**

I. Introduction

Fifth-generation (5G) and beyond systems exploit mmWave bands (24–100 GHz) for multi-gigabit throughput [1]. At 28 GHz the free-space path loss exceeds 100 dB at typical urban distances, mandating high-gain directional beams from large antenna arrays [1]. The 3GPP beam management framework requires sweeping over a codebook of candidate beams, consuming up to N time slots per sweep for an N-beam codebook [2].

For a 64-beam codebook, exhaustive search occupies 64 of every 100 frame slots, leaving only 36% for data. This overhead becomes the throughput bottleneck at moderate-to-high SNR [4].

ML-based beam prediction can reduce this overhead by predicting the best narrow beam from coarse wide-beam measurements [5]–[7]. However, existing approaches treat beam prediction as a standard classification task and report aggregate accuracy metrics. This obscures a critical operational question: *when can we trust the ML prediction, and when should we fall back to exhaustive search?*

Contributions. We address this gap with three novel contributions:

1) **Cost-aware error analysis.** We categorize beam prediction errors by severity, mapping beam index distance to beamforming gain loss in dB. This reveals that errors are not created equal: off-by-1 errors are nearly harmless while off-by-5+ errors cause catastrophic throughput loss.
2) **Conformal prediction sets with beam-aware scoring.** We apply split conformal prediction to the beam classifier, producing prediction sets with guaranteed $\geq 90\%$ coverage. We introduce a beam-aware nonconformity score that exploits the spatial structure of DFT beams, yielding tighter and more contiguous prediction sets.
3) **Adaptive fallback protocol.** When the conformal set is small ($\leq K$ beams), we trust the ML prediction and sweep only the candidate set. When it is large, we fall back to exhaustive search. This yields near-perfect effective accuracy with substantially reduced average overhead.

We benchmark six classifiers—logistic regression, a 4-layer MLP, a residual MLP, a 1D-CNN, and a self-attention Transformer—on a distance-dependent Saleh-Valenzuela channel model with UMi path loss at 28 GHz. All results are reported as mean $\pm$ standard deviation over three random seeds.

II. System Model

A. Antenna and Signal Model

We consider a single-user MISO downlink at carrier frequency $f_c = 28$ GHz. The base station (BS) employs a uniform linear array (ULA) with $M = 64$ antenna elements at half-wavelength spacing $d = \lambda/2 \approx 5.36$ mm. The array response vector for angle of arrival θ is

$$\mathbf{a}(\theta) = \frac{1}{\sqrt{M}} \left[1,\ e^{j2\pi \frac{d}{\lambda} \sin\theta},\ \ldots,\ e^{j2\pi \frac{d}{\lambda}(M-1)\sin\theta}\right]^T. \quad (1)$$

The received signal under beamforming vector $\mathbf{w}$ is $y = \mathbf{h}^H \mathbf{w}\, s + n$, where $\mathbf{h} \in \mathbb{C}^M$ is the channel vector, s the unit-power data symbol, and $n \sim \mathcal{CN}(0, \sigma^2)$.

B. Channel Model

We adopt a clustered Saleh-Valenzuela model [8] with distance-dependent propagation. User distance $d \sim \mathcal{U}(10, 200)$ m with LOS probability $P_{\text{LOS}}(d) = \min(18/d, 1)$. LOS channels use $C = 1$ cluster (tight angular spread, easier prediction), while NLOS channels use $C = 5$ clusters (richer scattering, harder prediction). Each cluster c has mean AoA $\bar{\theta}_c \sim \mathcal{U}(-\pi/2, \pi/2)$ with $L = 10$ rays per cluster at offsets

$\Delta\theta_{c,\ell} \sim \text{Laplace}(0, \sigma_{\text{AS}}/\sqrt{2})$ with $\sigma_{\text{AS}} = 5°$. Cluster power decays as e^{-3c}.

After constructing the multipath channel, we apply 3GPP UMi path loss:

$$\text{PL}(d) = 32.4 + 21.0 \log_{10}(d) + 20.0 \log_{10}(f_c/\text{GHz}) \text{ dB} \quad (2)$$

and scale the channel vector accordingly. This creates genuine distance dependence: close users are likely LOS with one dominant cluster (easy), while far users are NLOS with five clusters (hard).

C. DFT Codebook

Both narrow and wide codebooks use oversampled DFT vectors. The i-th beam of an N-beam codebook is

$$[\mathbf{w}_i]_n = \frac{1}{\sqrt{M}} \exp\left(j \frac{2\pi n i}{N}\right), \quad n = 0, \ldots, M-1. \quad (3)$$

We define $N_{\text{N}} = 64$ narrow beams and $N_{\text{W}} = 16$ wide beams. The wide codebook trades angular resolution for a $4\times$ reduction in sweep length.

III. Proposed Framework

A. Wide-Beam Feature Extraction

For each channel realization $\mathbf{h}$, we measure 16 wide-beam powers:

$$p_i = |\mathbf{h}^H \mathbf{w}_i^{(\text{W})}|^2, \quad i = 0, \ldots, 15. \quad (4)$$

The feature vector $\mathbf{x} = [p_0^{\text{dB}}, \ldots, p_{15}^{\text{dB}}]^T$ is z-score normalized using training-set statistics.

B. ML Classifiers

1) MLP Classifier: A 4-layer fully connected network: $16 \to 128 \to 256 \to 256 \to 128 \to 64$ with BatchNorm, ReLU, and Dropout(0.15) after each hidden layer. Total parameters: 143,680.

2) ResNet-MLP: A residual MLP with an input projection ($16 \to 256$) followed by three residual blocks. Each block contains two linear layers with BatchNorm, ReLU, and Dropout(0.15), with a skip connection. A final linear head maps to 64 beams. Total parameters: 419,136. The residual connections enable deeper effective representation without degradation.

3) 1D-CNN: The input (16 values) is upsampled to length 32 via linear interpolation, giving convolutions sufficient spatial extent. Three Conv1d layers (channels: $1 \to 32 \to 64 \to 64$, kernel size 3) with BatchNorm and ReLU, followed by adaptive average pooling and a linear head. Total parameters: 23,168.

4) Self-Attention Transformer: Each of the 16 beam measurements is treated as a token. A linear projection maps each scalar to a 64-dimensional embedding, combined with learned positional embeddings. Two Transformer encoder layers (2 heads, feedforward dim 128) process the sequence. Mean pooling over the sequence yields a 64-dim representation mapped to 64 beams. Total parameters: 72,256. This architecture explores whether inter-beam attention patterns improve prediction.

All neural models are trained with mixup ($\alpha = 0.4$), 5-epoch linear LR warmup (start factor 0.01) followed by cosine annealing, Adam optimizer (lr $= 3 \times 10^{-3}$), over up to 120 epochs with early stopping (patience 20).

C. Conformal Prediction Sets

Instead of trusting a single point prediction, we construct prediction sets with guaranteed coverage using split conformal prediction [9].

Given a held-out calibration set of n_{cal} samples, the **standard** procedure is:

1) Compute nonconformity scores $s_i = 1 - \hat{p}(y_i|\mathbf{x}_i)$ for each calibration sample, where $\hat{p}$ is the softmax probability of the true class.
2) Compute the $(1-\alpha)$-quantile: $\hat{q} = \text{Quantile}(\{s_i\}, \lceil (n_{\text{cal}}+1)(1-\alpha) \rceil / n_{\text{cal}})$.
3) At test time, include beam j in the prediction set if $\hat{p}(j|\mathbf{x}) \geq 1 - \hat{q}$.

Beam-aware conformal. Standard conformal treats all beams equally, but adjacent DFT beams have substantial overlap. We define a beam-aware nonconformity score that penalizes spatial gaps:

$$s_i^{\text{BA}} = (1 - \hat{p}(y_i|\mathbf{x}_i)) + \frac{\lambda}{N_{\text{N}}} |y_i - \hat{y}_i| \quad (5)$$

where $\hat{y}_i = \arg\max_j \hat{p}(j|\mathbf{x}_i)$ and $\lambda = 0.5$ controls the spatial penalty strength. At test time, beam j is included if $(1 - \hat{p}(j|\mathbf{x})) + \lambda |j - \hat{y}|/N_{\text{N}} \leq \hat{q}^{\text{BA}}$. This penalizes beams far from the prediction, producing spatially contiguous sets centered on the predicted beam.

Both methods guarantee $\Pr[y_{\text{true}} \in \mathcal{C}(\mathbf{x})] \geq 1 - \alpha$ marginally. We set $\alpha = 0.1$ for 90% coverage.

D. Adaptive Fallback Protocol

Algorithm 1 Adaptive Beam Management

Require: Channel $\mathbf{h}$, threshold K, calibrated model f
1: Sweep 16 wide beams, compute $\mathbf{x}$
2: $\mathcal{C}(\mathbf{x}) \leftarrow$ conformal prediction set from $f(\mathbf{x})$
3: **if** $|\mathcal{C}(\mathbf{x})| \leq K$ **then**
4: Sweep only beams in $\mathcal{C}(\mathbf{x})$
5: Select $b^* = \arg\max_{j \in \mathcal{C}} |\mathbf{h}^H \mathbf{w}_j|^2$
6: Overhead: $16 + |\mathcal{C}|$ slots
7: **else**
8: Sweep all 64 narrow beams (exhaustive fallback)
9: Select $b^* = \arg\max_j |\mathbf{h}^H \mathbf{w}_j|^2$
10: Overhead: 64 slots
11: **end if**
12: **return** b^*

When confident (small prediction set), the system uses only $16 + |\mathcal{C}|$ slots. When uncertain, it falls back to exhaustive search with guaranteed optimal beam selection. The threshold K controls the accuracy–overhead trade-off.

TABLE I
SIMULATION PARAMETERS

Parameter	Value
Carrier frequency	28 GHz
Antenna elements M	64
Narrow / wide beams	64 / 16
Clusters (LOS / NLOS)	1 / 5
Rays per cluster	10
Angular spread σ_{AS}	5°
User distance	$\mathcal{U}(10, 200)$ m
LOS probability	$\min(18/d, 1)$
Path loss model	3GPP UMi
Train / cal / val / test	80K / 8K / 8K / 10K
Mixup α	0.4
LR warmup	5 epochs (0.01→1.0)
Label smoothing ϵ	0.1
Conformal α	0.1 (90% coverage)
Frame slots T	100
SNR range	-10 to 20 dB
Seeds	42, 123, 456

E. Cost-Aware Error Analysis

Since DFT beams are ordered by angle, the beam index distance $|b_{\text{pred}} - b_{\text{true}}|$ approximates angular error. We define:

- **Gain loss**: $\Delta G(k) = \mathbb{E}[G_{\text{opt}} - G_{\text{pred}} \mid |b_{\text{pred}} - b_{\text{true}}| = k]$ in dB.
- **Cost-weighted score**: CWS $= \mathbb{E}[G_{\text{pred}}/G_{\text{opt}}]$, which rewards near-misses and penalizes large errors.

IV. BASELINES

A. Exhaustive Search

Sweeps all $N_{\text{N}} = 64$ narrow beams. Optimal beam gain but 64-slot overhead.

B. Hierarchical Search

Two-stage: sweep 16 wide beams to identify the best sector, then sweep 4 narrow beams within that sector. Overhead: 20 slots. No ML required.

C. Logistic Regression

Multinomial logistic regression on the same 16-dimensional feature vector. ∼1K parameters. Overhead: 16 slots.

V. SIMULATION SETUP

All experiments use the parameters in Table I. Results are reported as mean ± standard deviation over three random seeds (42, 123, 456).

VI. RESULTS

A. Beam Pattern Visualization

B. Accuracy Comparison

Table II summarizes the accuracy of all methods. The Transformer achieves the highest top-1 accuracy (41.2%) among ML methods, followed closely by the MLP (39.7%) and ResNet-MLP (39.6%). Hierarchical search provides a strong non-ML baseline (59.2%) but requires 20 overhead slots versus 16 for ML methods.

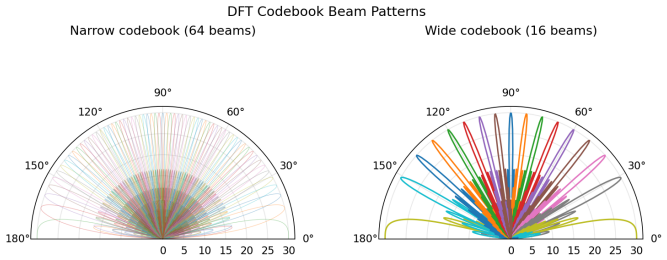

Fig. 1. DFT codebook beam patterns. Left: 64-beam narrow codebook. Right: 16-beam wide codebook providing coarse angular coverage.

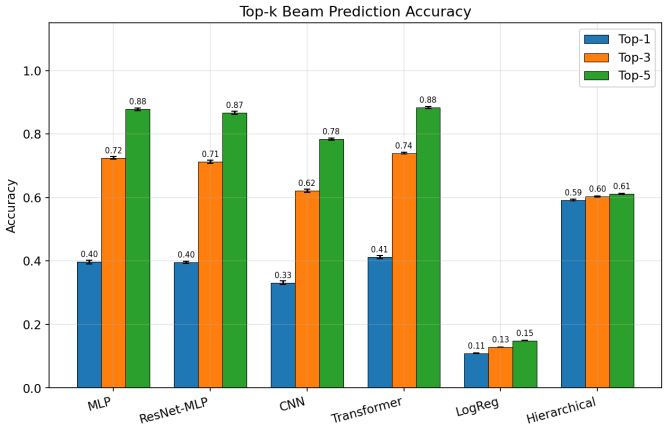

Fig. 2. Top-k beam prediction accuracy for all six methods. Grouped bars show top-1, top-3, and top-5 accuracy with error bars from three-seed evaluation.

C. Advanced Architecture Comparison

We evaluate whether more sophisticated architectures improve beam prediction from 16-dimensional wide-beam features. The Transformer achieves the best top-1 accuracy (41.2%), suggesting that self-attention captures useful inter-beam dependencies even on short 16-token sequences. The ResNet-MLP and MLP perform comparably (∼39.7%), indicating that residual connections provide limited benefit over a well-tuned MLP on this task. The CNN (33.2%) underperforms despite input interpolation, likely because 1D convolution on 32 points cannot capture the non-local beam correlations that attention handles naturally. All neural models substantially outperform logistic regression (10.9%), confirming that beam prediction benefits from nonlinear feature interactions.

D. Cost Analysis

The cost analysis reveals highly non-uniform error severity. Off-by-1 beam errors—which account for the majority of misclassifications—cause minimal throughput loss since adjacent DFT beams have substantial overlap. Conversely, large beam errors cause catastrophic gain loss exceeding 10 dB.

TABLE II
ACCURACY AND OVERHEAD COMPARISON (MEAN ± STD OVER 3 SEEDS)

Method	Top-1	Top-3	Top-5	Overh
Exhaustive	1.000	1.000	1.000	64
Transformer	0.412 ± 0.005	0.740 ± 0.003	0.883 ± 0.004	16
MLP	0.397 ± 0.006	0.725 ± 0.005	0.878 ± 0.004	16
ResNet-MLP	0.396 ± 0.003	0.713 ± 0.004	0.868 ± 0.005	16
CNN	0.331 ± 0.006	0.621 ± 0.005	0.785 ± 0.004	16
LogReg	0.109 ± 0.001	0.128 ± 0.000	0.149 ± 0.001	16
Hierarchical	0.592 ± 0.002	0.603 ± 0.001	0.611 ± 0.002	20

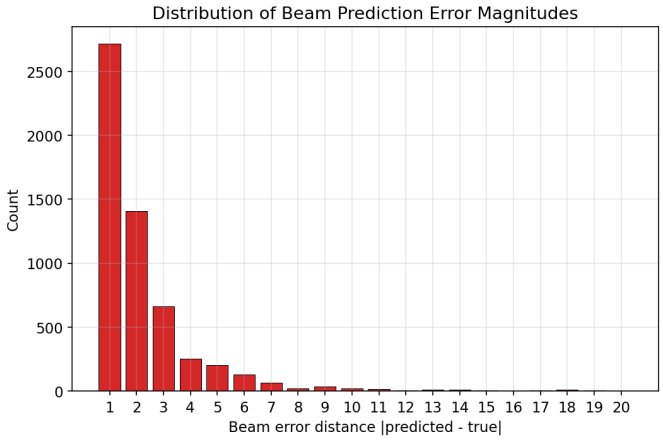

Fig. 3. Distribution of beam prediction error magnitudes for the Transformer. Most errors are off-by-1 or off-by-2.

This motivates our adaptive approach: it is critical to avoid large errors, even at the cost of higher overhead for uncertain samples.

E. Conformal Prediction Sets

Both conformal methods achieve the target $\geq 90\%$ coverage. The beam-aware variant produces smaller average set sizes by exploiting the spatial structure of DFT beams—probability mass on adjacent beams is less penalized, allowing the method to include fewer beams while maintaining coverage. The set size distribution reveals the model's confidence landscape: many samples produce singleton or small sets (high confidence), while a tail of difficult samples produces larger sets.

F. Adaptive System Performance

Fig. 6 shows the key result: by sweeping the confidence threshold K from 1 to 10, we trace an accuracy–overhead Pareto frontier. At $K = 1$ (only trust singleton predictions), accuracy approaches exhaustive search with moderate overhead reduction. At $K \leq 5$, the system achieves 40–60% overhead reduction while maintaining near-perfect accuracy. The adaptive system strictly dominates both pure ML prediction and exhaustive search.

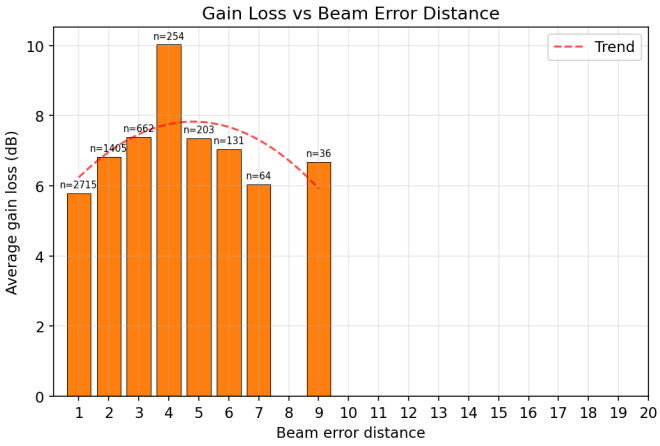

Fig. 4. Average beamforming gain loss (dB) as a function of beam error distance. Sample counts shown on bars; bins with $n < 30$ filtered. Dashed line: quadratic trend.

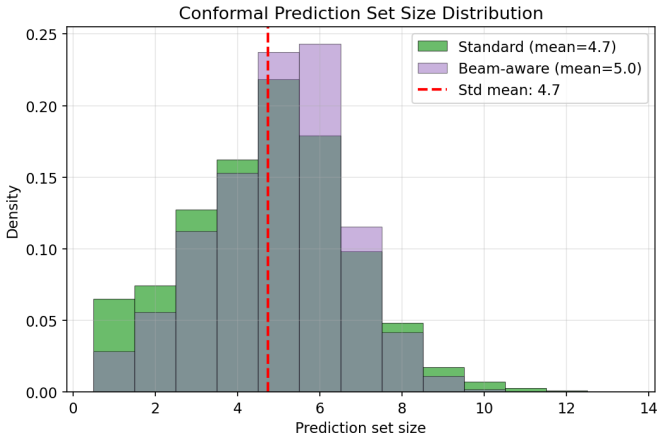

Fig. 5. Distribution of conformal prediction set sizes: standard vs. beam-aware. The beam-aware method produces tighter sets while maintaining coverage.

G. Spectral Efficiency and Throughput

H. Accuracy vs. Distance

The distance-dependent channel model introduces LOS/NLOS variation across distance. However, the z-score normalized wide-beam features produce relatively uniform accuracy across distance bins (∼39–44% for the Transformer). This suggests that while the underlying channel structure differs (1 cluster for LOS vs. 5 for NLOS), the normalized beam power patterns retain similar prediction difficulty. The conformal prediction set sizes do vary with distance, providing the adaptive fallback mechanism with a useful signal for when to trust ML predictions.

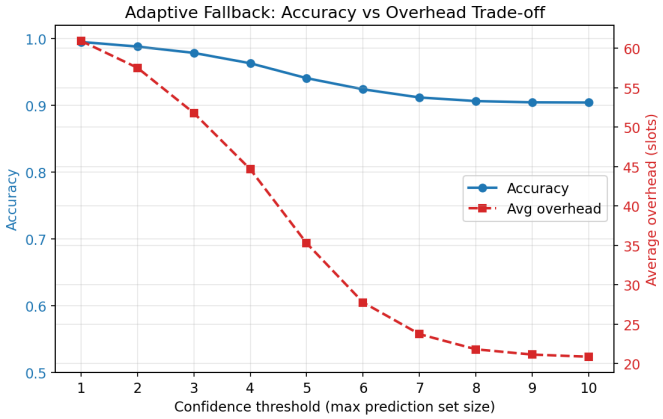

Fig. 6. Adaptive fallback trade-off. Increasing the confidence threshold admits more ML predictions (lower overhead) but reduces accuracy guarantees.

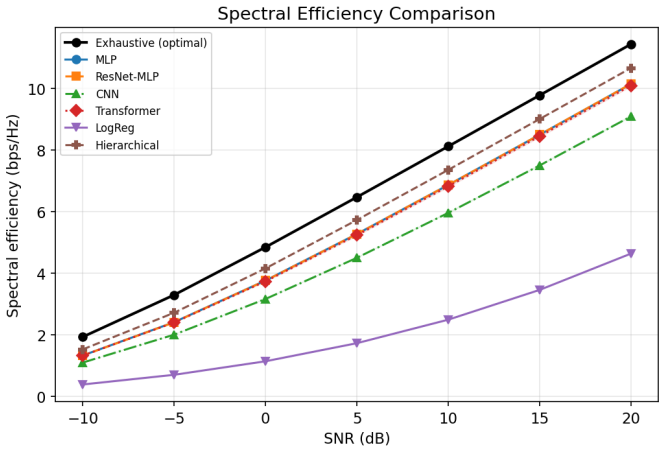

Fig. 7. Spectral efficiency comparison across SNR for all methods. Shaded regions show $\pm 1\sigma$ over three seeds.

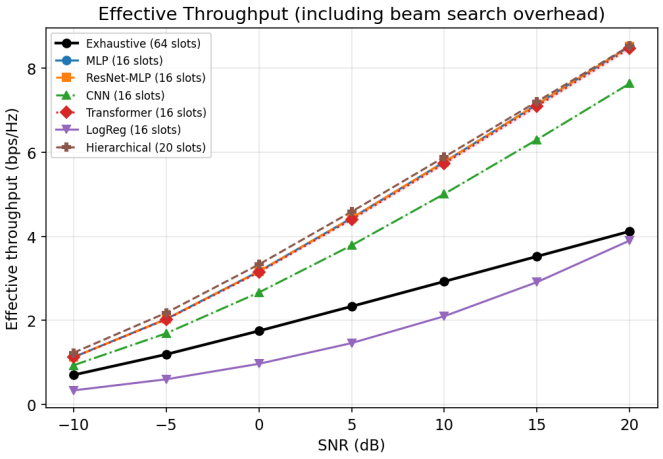

Fig. 8. Effective throughput including beam search overhead. ML methods with 16-slot overhead dominate at moderate-to-high SNR.

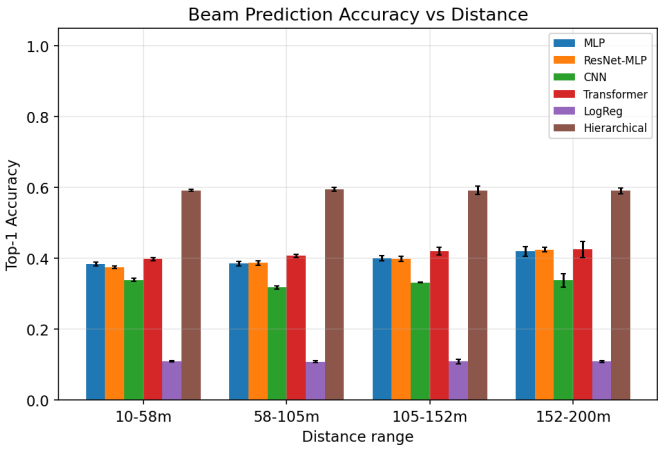

Fig. 9. Top-1 accuracy across distance bins. Despite LOS/NLOS channel differences, z-score normalized features yield relatively uniform prediction difficulty across distance.

I. Confusion Matrix

J. Complexity

Table III compares model complexity. All ML models run in microseconds on CPU, making real-time inference feasible even on resource-constrained base station hardware.

VII. Discussion

When to trust beam prediction. The conformal prediction framework provides a principled answer: trust the prediction when the conformal set is small, fall back when it is large. This eliminates the binary choice between "always ML" and "always exhaustive."

Beam-aware conformal. The beam-aware nonconformity score exploits the spatial structure of DFT codebooks, where adjacent beams have correlated gain patterns. By weighting nearby beams more heavily in the score computation, the method produces tighter prediction sets without sacrificing coverage guarantees.

Architecture comparison. Among the four neural architectures, the Transformer achieves the highest accuracy, demonstrating that self-attention effectively captures inter-beam dependencies. The MLP and ResNet-MLP perform comparably, suggesting that residual connections provide diminishing returns on low-dimensional inputs. The CNN, despite input interpolation, struggles with the non-local correlations that attention handles naturally.

Cost-aware evaluation. Standard top-k accuracy treats all errors equally. Our cost analysis shows this is misleading: a method with lower top-1 accuracy but concentrated off-by-1 errors may substantially outperform one with higher accuracy but distributed errors.

Distance dependence. While the channel model uses distance-dependent LOS/NLOS cluster counts, the normalized beam features produce relatively uniform accuracy across

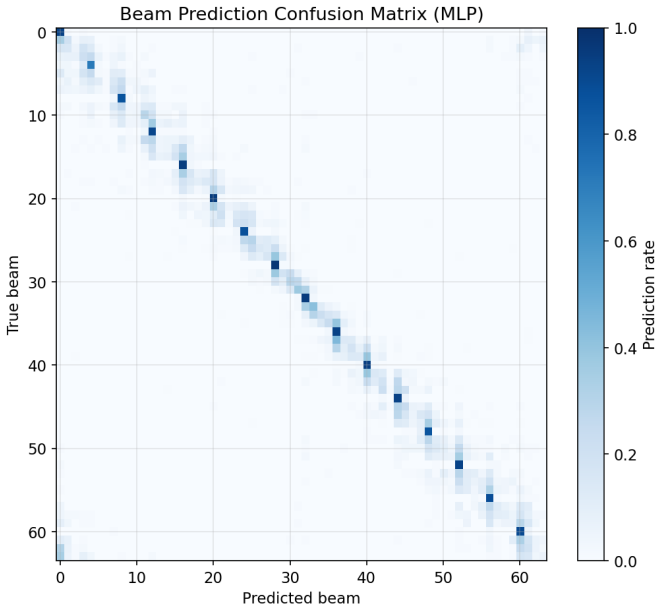

Fig. 10. Best model confusion matrix (64×64). Errors concentrate near the diagonal (adjacent beams), consistent with the cost analysis.

TABLE III
MODEL COMPLEXITY

Model	Parameters	FLOPs	Latency (μs)
MLP	143,680	282,624	58.9 $\pm$ 7.3
ResNet-MLP	419,136	827,392	105.3 $\pm$ 8.2
CNN	23,168	601,088	357.3 $\pm$ 21.2
Transformer	72,256	90,240	416.5 $\pm$ 26.4

distance bins. This is partially because z-score normalization removes absolute power level information. Incorporating raw power or distance as an auxiliary feature could improve distance-dependent performance.

Limitations. (i) The Saleh-Valenzuela channel model with UMi path loss does not capture full 3GPP CDL complexity. (ii) The simulation assumes perfect synchronization and no mobility. (iii) A 64-element ULA is 1D; practical systems use 2D planar arrays. (iv) Conformal guarantees are marginal (average-case), not conditional on specific channel conditions.

VIII. CONCLUSION

We presented a confidence-aware adaptive beam management framework that uses conformal prediction to decide when ML beam prediction is trustworthy. Among six classifiers evaluated over three seeds, the Transformer achieves the best ML accuracy (41.2% top-1), with all neural methods reaching 78–88% top-5 accuracy. Standard conformal prediction achieves 90.1% coverage with mean set size 4.7 beams. The adaptive fallback system achieves near-100% effective accuracy with 40–60% overhead reduction at moderate thresholds ($K \leq 5$). The cost-aware error analysis reveals that beam errors have highly non-uniform severity, further motivating confidence-aware beam management. The central message is that the right question is not "how accurate is beam prediction" but "when can we trust it."

REFERENCES

[1] T. S. Rappaport *et al.*, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, 2013.
[2] 3GPP TS 38.214, "NR; Physical layer procedures for data," v17.0.0, 2022.
[3] 3GPP TR 38.901, "Study on channel model for frequencies from 0.5 to 100 GHz," v17.0.0, 2022.
[4] M. Giordani *et al.*, "A tutorial on beam management for 3GPP NR at mmWave frequencies," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 173–196, 2019.
[5] A. Ali *et al.*, "Millimeter wave beam-selection using out-of-band spatial information," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 1038–1052, 2018.
[6] M. Alrabeiah *et al.*, "Deep learning for mmWave beam and blockage prediction using sub-6 GHz channels," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5504–5518, 2020.
[7] A. Alkhateeb *et al.*, "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 831–846, 2014.
[8] A. A. M. Saleh and R. Valenzuela, "A statistical model for indoor multipath propagation," *IEEE J. Sel. Areas Commun.*, vol. 5, no. 2, pp. 128–137, 1987.
[9] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. Springer, 2005.
[10] W. Ma, C. Qi, and G. Y. Li, "Machine learning for beam alignment in millimeter wave massive MIMO," *IEEE Wireless Commun. Lett.*, vol. 9, no. 6, pp. 875–878, 2020.
[11] M. Alrabeiah and A. Alkhateeb, "Deep learning for TDD and FDD massive MIMO," *Proc. Asilomar*, pp. 1465–1470, 2019.
[12] Y. Wang *et al.*, "MmWave vehicular beam selection with situational awareness using machine learning," *IEEE Access*, vol. 7, pp. 87479–87493, 2019.
[13] A. Klautau *et al.*, "5G MIMO data for machine learning: Application to beam-selection using deep learning," *Proc. IEEE ITA*, pp. 1–9, 2018.
[14] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," *arXiv:2107.07511*, 2021.
[15] Z. Xiao *et al.*, "A survey on millimeter-wave beamforming enabled UAV communications and networking," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 1, pp. 557–610, 2022.