4th Workshop on Uncertainty Quantification for Computer Vision

CVPR 2025 Workshop
Wednesday, 11th June 2025, Full day
Room 102 B

WQLCP: Weighted Adaptive Conformal Prediction for Robust Uncertainty Quantification Under Distribution Shifts

Shadi Alijani, Homayoun Najjaran
University of Victoria, BC, Canada

4th workshop in Uncertainty Quantification for Computer Vision

CVPR 2025

Introduction and Motivation

01 In safety-critical applications, high accuracy alone is insufficient

02 Understanding and quantifying uncertainty is crucial for reliable and trustworthy predictions

03 The ability to interpret model predictions and quantify uncertainty enhance both algorithmic robustness and real-world applicability

04 Conformal prediction provides a framework for constructing prediction sets with guaranteed coverage, assuming **exchangeable data**

05 Real-world scenarios often involve distribution shifts that violate exchangeability assumptions about train and test data

BACKGROUND

How does conformal prediction work?

01

Conformal Prediction (CP) provides a framework for constructing prediction sets with guaranteed coverage, Assumes exchangeable data (i.i.d. assumption)

02

For a dataset D, CP constructs a set-valued function that ensures:

$$P_{(x,y)\sim D}(y \in \mathcal{C}_D(x)) \geq 1-\alpha.$$

03

It Generates prediction set:

$$\mathcal{C}_D(x;\tau_D) := \{y' \in \mathcal{Y} : s(x,y') \geq \tau_D\}.$$

04

Threshold τ is chosen as the α(1 + 1/n)-quantile of empirical scores $\{s(x_i,y_i)\}_{i=1}^{n}$

PROBLEM STATEMENT

- Distribution shifts violate exchangeability assumption
- Leads to unreliable coverage and inflated prediction sets
- Existing methods rely on simplifying assumptions about shifts
- Many use simplistic nonconformity metrics insufficient for high-dimensional image data

01

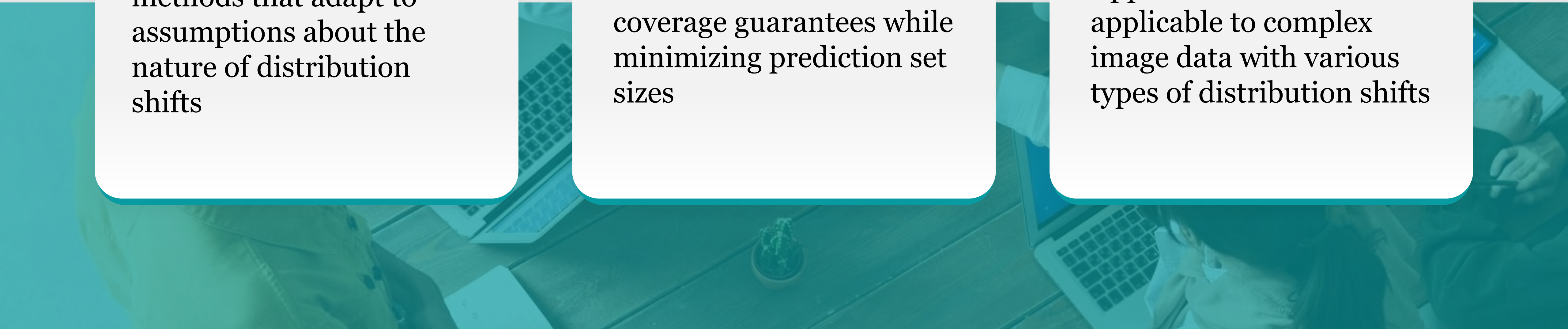
Need for adaptive methods that adapt to assumptions about the nature of distribution shifts

02

Methods should maintain coverage guarantees while minimizing prediction set sizes

03

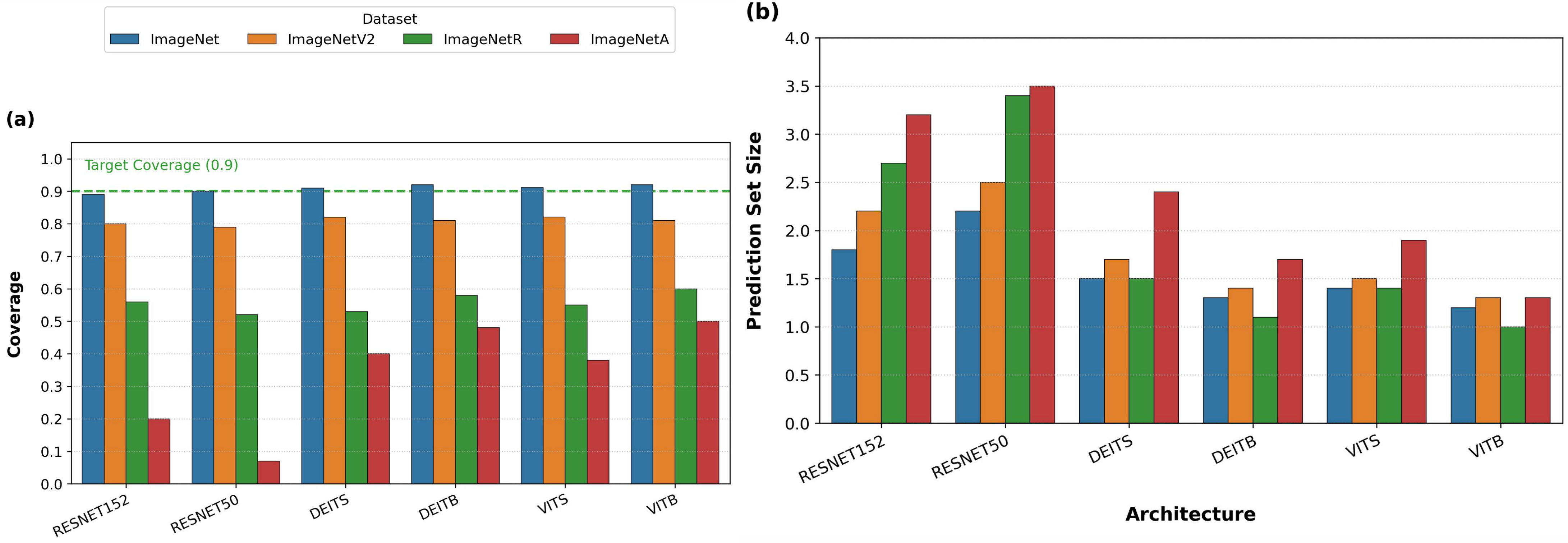
Approach should be applicable to complex image data with various types of distribution shifts



Coverage/ Set Size Violation Under Distribution Shifts

First Experiment:

• CP coverage violations under distribution shifts • Transformers (ViTs, DeiTs) outperform CNNs (ResNet50/152) • All models fail to maintain 0.90 target coverage under shifts • Prediction set sizes increase for shifted datasets • ImageNet variants require larger prediction sets



Proposed Method

RLSCP (Reconstruction Loss-Scaled Conformal Prediction)

$$\mathcal{L}_{\text{test}} = \{\mathcal{L}(x_1), \ldots, \mathcal{L}(x_N)\}$$

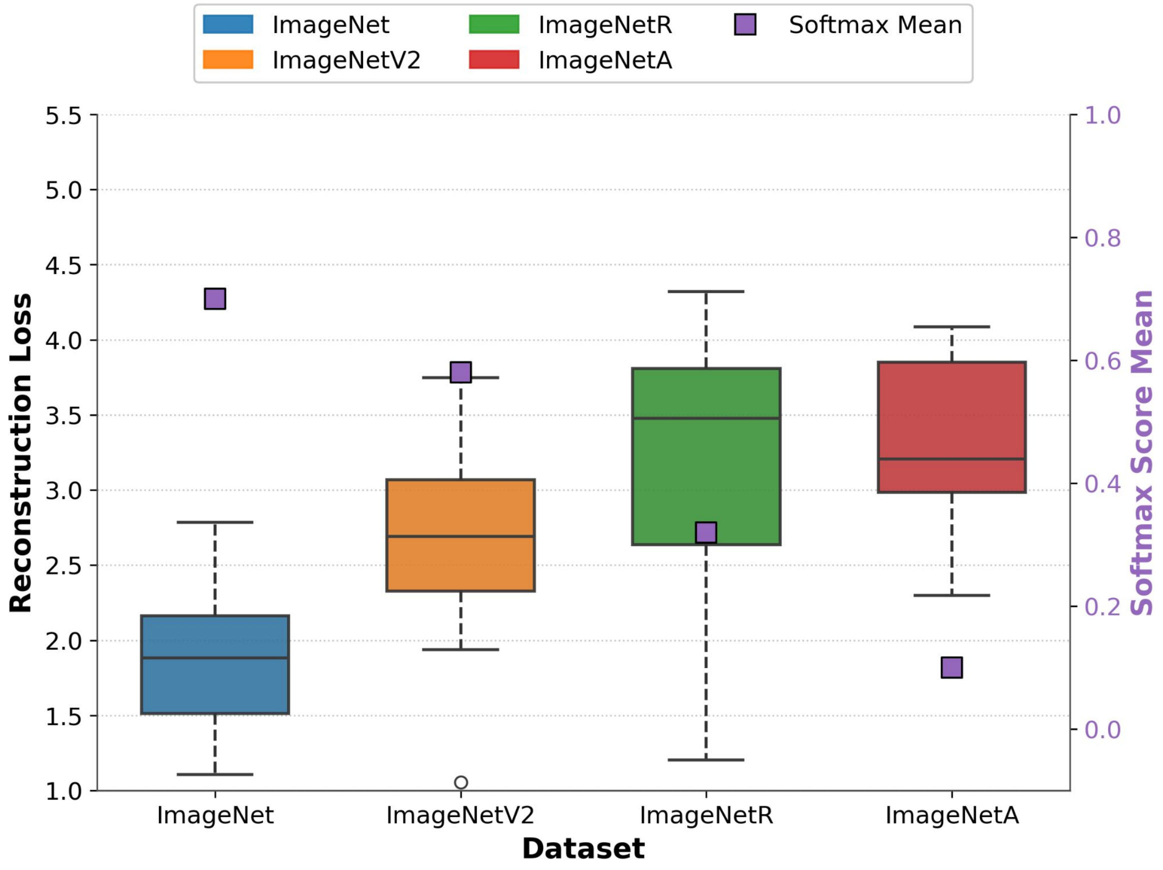
$$RL_{\text{test}} = q_{1-\alpha}(\mathcal{L}_{\text{test}})$$

$$C_{\text{test}}(x; \tau_{\text{test}}) = \{y' \in \mathcal{Y} : s(x, y') \cdot \max(1, RL_{\text{test}}) \geq \tau_{\text{test}}\}$$

- Utilizes reconstruction losses from a Variational Autoencoder (VAE) as an uncertainty metric
- Scales prediction set sizes using VAE-derived reconstruction losses
- Links reconstruction uncertainty to conformal scores

Compatibility: Preserves the original SplitCP prediction sets when RLtest ≤ 1

Adaptivity: Scales the score function proportionally to reconstruction loss when RLtest > 1



Why Reconstruction Loss as an Uncertainty Metric?

Strong correlation between VAE reconstruction loss and performance degradation under distribution shift

Algorithm 1 Weighted Quantile Loss-Scaled Conformal Prediction (WQLCP)

Require: Test dataset $\mathcal{D}_{\text{test}} = \{x_i\}_{i=1}^{n}$, Calibration dataset $\mathcal{D}_{\text{cal}} = \{x_j\}_{j=1}^{m}$, Confidence level $1 - \alpha$.

Ensure: Prediction sets $\{\mathcal{C}(x_i)\}_{i=1}^{n}$.

Step 1: Compute Reconstruction Losses

for $x_j \in \mathcal{D}_{\text{cal}}$ **do** Compute $L_{\text{cal}}(x_j; \theta)$ using VAE.

end for

for $x_i \in \mathcal{D}_{\text{test}}$ **do** Compute $L_{\text{test}}(x_i; \theta)$ using VAE.

end for

Step 2: Calculate Weights

for $x_j \in \mathcal{D}_{\text{cal}}$ **do** Compute $w(x_j) \propto \frac{L_{\text{cal}}(x_j;\theta)}{L_{\text{test}}(x_i;\theta)+\epsilon}$.

end for

Step 3: Compute Weighted Quantile Threshold

$\hat{q} = \inf\left\{q : \sum_{j=1}^{n} w(x_j)\mathbb{I}\{s_j \leq q\} \geq (1-\alpha)\sum_{j=1}^{n} w(x_j)\right\}$

Step 4: Scale Test Scores

for $x_i \in \mathcal{D}_{\text{test}}$ **do** Compute $s_{\text{scaled}}(x_i, y) = s(x_i, y) \cdot \max(1, RL_{\text{test}})$.

end for

Step 5: Generate Prediction Sets

for $x_i \in \mathcal{D}_{\text{test}}$ **do** Construct $\mathcal{C}(x_i) = \{y : s_{\text{scaled}}(x_i, y) \geq \tau_{\text{test}}\}$.

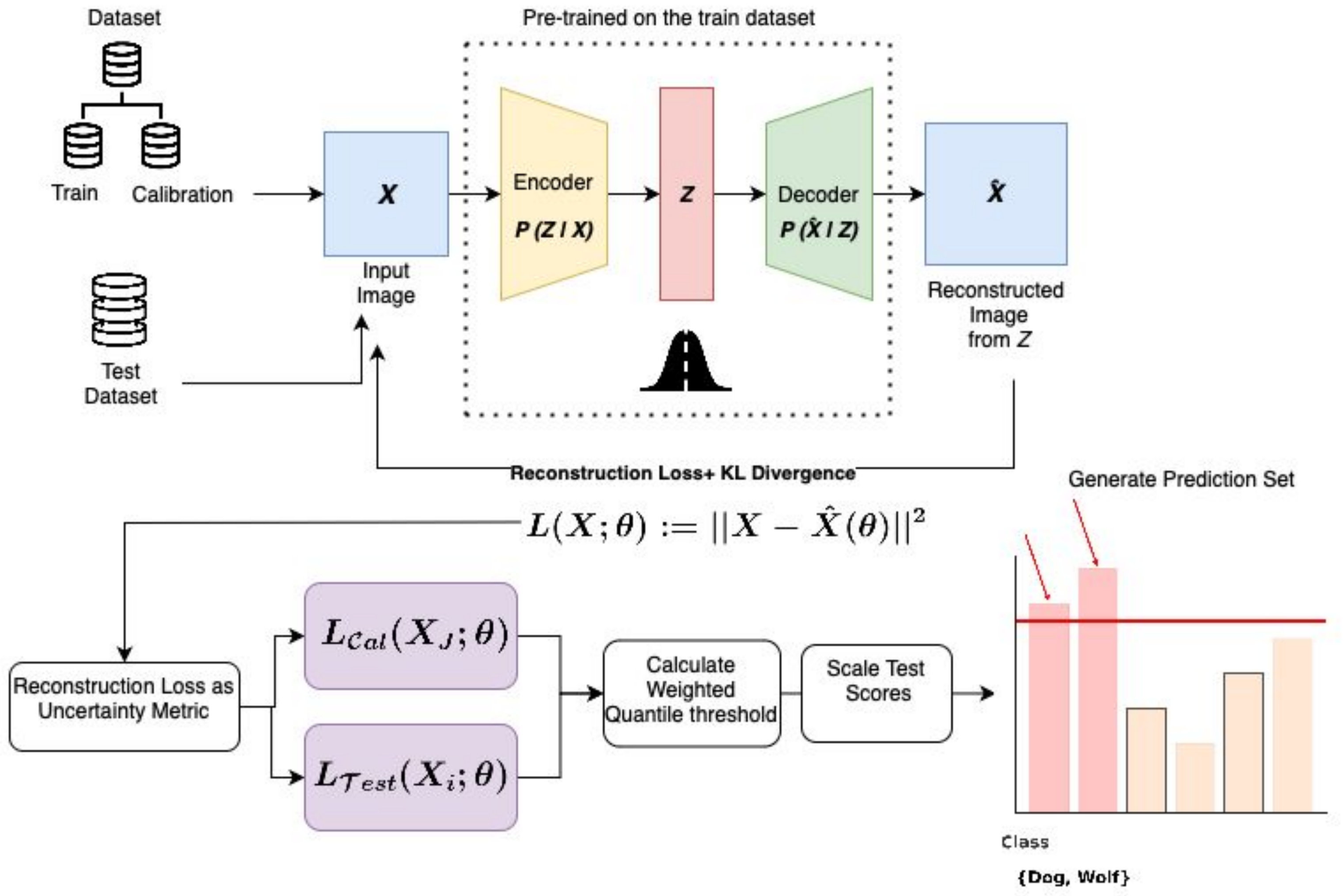
end for

Step 6: Return Prediction Sets Return $\{\mathcal{C}(x_i)\}$.

WQLCP (weighted Quantile Loss-Scaled Conformal Prediction)

WQLCP Framework

The model leverages VAEs to compute reconstruction losses for uncertainty estimation, applying weighted quantile and scaling test scores to refine conformal prediction



Experimental Setup

Datasets: ImageNet (baseline)

ImageNetV2: 10,000 images with natural distribution shifts (mild shift)

ImageNetR: 30,000 artistic renditions of 200 ImageNet classes (domain gap)

ImageNetA: 7,500 adversarially filtered natural images (extreme shift)

Model Architectures:

CNN: ResNet-50, ResNet-152

Transformer-based: ViT-S/B, DeiT-S/B

Metrics:

Coverage

Prediction Set Size

Shift Severity

Training:

100 epochs,

optimizer: AdamW

lr=10^-4, Batch size=256

Quantitative Performance

Method	Backbone	ImageNet	ImageNetV2	ImageNetR	ImageNetA
Baselines					
Naive	RN50	0.7523 / 4.5123	0.6462 / 4.4837	0.3521 / 5.3459	0.0152 / 5.5342
	RN152	0.7525 / 4.2321	0.6546 / 4.1487	0.3869 / 4.8132	0.1211 / 5.1028
	ViT-S	0.7721 / 3.8528	0.6689 / 3.7342	0.4034 / 4.3124	0.3568 / 4.7341
	ViT-B	0.7723 / 3.4457	0.6732 / 3.3759	0.4521 / 3.9123	0.3818 / 4.1128
	DeiT-S	0.7625 / 3.2124	0.6834 / 3.1187	0.3741 / 3.6348	0.3021 / 3.9234
Proposed Methods					
RLSCP (Ours)	RN-VAE	0.9201 / 2.5001	0.9102 / 7.8003	0.7503 / 21.002	0.6002 / 50.512
	ViT-VAE	0.9301 / 2.2003	0.9304 / 7.0002	0.7992 / 16.703	0.7703 / 28.514
WQLCP (Ours)	RN-VAE	0.9402 / 2.0002	0.9292 / 8.9001	0.7773 / 11.112	0.6501 / 12.122
	ViT-VAE	**0.9503 / 1.8001**	**0.9403 / 6.7003**	**0.8502 / 8.302**	**0.7402 / 9.501**
Comparison Against SOTA					
WCP [41]	—	0.8801 / 10.5002	0.8702 / 18.8002	0.7503 / 8.5001	0.6503 / 7.3002
SSCP [37]	—	0.9001 / 11.2002	0.8603 / 17.0003	0.7703 / 9.7001	0.6702 / 8.6003
RAPS [3]	RN50	0.9002 / 3.7821	0.7671 / 2.2521	0.4823 / 3.4923	0.0221 / 3.0821
	RN152	0.9021 / 2.9821	0.7823 / 2.0123	0.5223 / 3.1421	0.1723 / 3.0523
	ViT-S	0.9023 / 1.7223	0.8323 / 2.0923	0.5923 / 3.5921	0.4723 / 3.2923
	ViT-B	0.9021 / 1.5423	0.8423 / 1.8923	0.6721 / 3.2023	0.5321 / 3.1321
	DeiT-S	0.8999 / 2.0921	0.8421 / 2.6421	0.5923 / 5.2521	0.5099 / 5.1923
	DeiT-B	0.8991 / 1.5921	0.8323 / 1.8923	0.6223 / 3.6623	0.4809 / 3.2723

Technical Insights

01 Strong correlation between VAE reconstruction loss and performance degradation under distribution shift

02 Transformer-based VAEs (ViT-VAE) outperform CNN-based VAEs (RN-VAE)

03 On ImageNetA (severe shift):
RN-VAE: Coverage drop of $\Delta = 0.290$
ViT-VAE: Smaller drop of $\Delta = 0.210$

04 WQLCP with ViT-VAE achieves 74.0% coverage with efficient set size of 9.50

05 WQLCP adaptively balances coverage and efficiency

06 Transformer-based VAEs better capture shift severity

β-VAE Ablation Study

β-VAE parameter optimization shows:

Optimal β = 1.2 achieves 0.7402 coverage on ImageNetA

Increasing β leads to higher KL divergence

(stronger regularization)

Reconstruction MSE decreases with higher β — indicating improved generalization

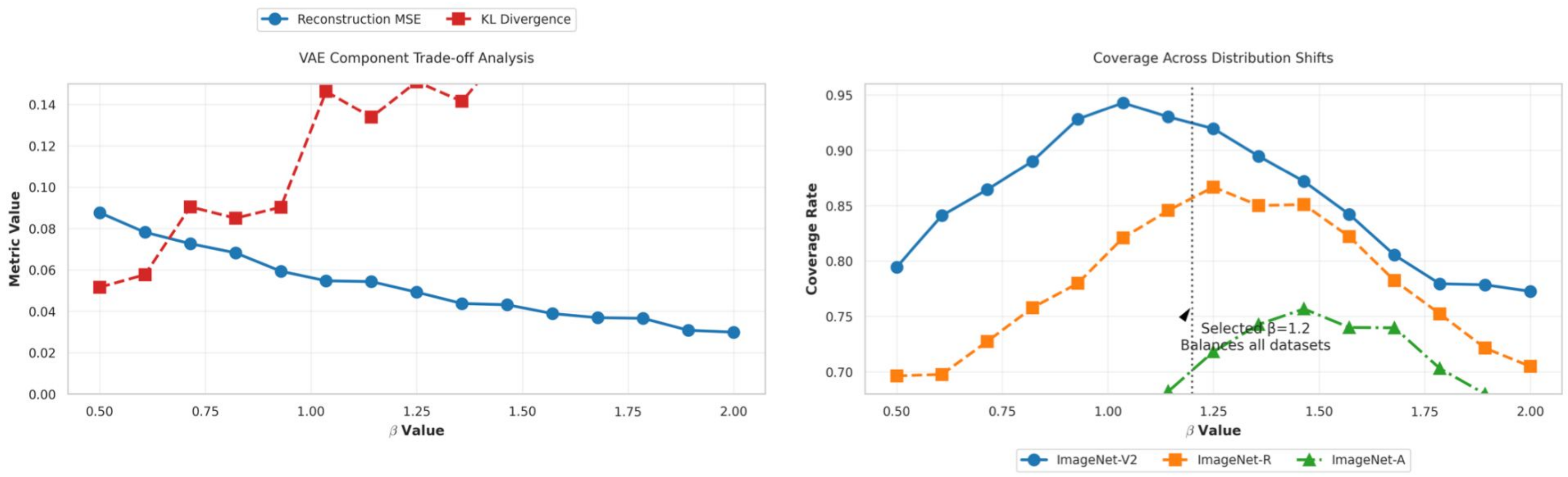


Figure 4. β-VAE ablation study: (Left) Reconstruction MSE loss vs KL divergence trade-off. (Right) Coverage vs β on ImageNetA, showing optimal $\beta = 1.2$ achieves 0.7402 coverage.

Failure Mode Analysis

Under-coverage in Extreme Shifts:

- 12.1% of ImageNetA samples with high reconstruction loss (Lrec > 3σ) exhibit severe under-coverage
- 0.42 coverage vs. average 0.74
- Linked to β-VAE over-regularization under adversarial shifts

Inflated Set Sizes in Fine-grained Classes:

- 4.8% of predictions require larger set sizes
- Occurs in visually similar classes (e.g., dog breeds)
- Suggests need for class-conditional calibration strategies
- These failure modes highlight opportunities for future research in adaptive regularization and class-conditional calibration techniques

CONCLUSIONS

Introduced RLSCP: Using VAE reconstruction losses to scale conformal scores

Developed WQLCP: Weighted quantile calibration framework for improved uncertainty estimation

Demonstrated state-of-the-art performance on ImageNet variants with distribution shifts

Provided a robust solution for conformal prediction under distribution shifts

THANK YOU

Q&A

Presenter: Shadi Alijani

Shadi Alijani, Homayoun Najjaran

University of Victoria, BC, Canada

Email: {shadialijani,najjaran}@uvic.ca