Understanding and Improving Adversarial Robustness of Neural Probabilistic Circuits

Weixin Chen
University of Illinois Urbana-Champaign
weixinc2@illinois.edu

Han Zhao
University of Illinois Urbana-Champaign
hanzhao@illinois.edu

Abstract

Neural Probabilistic Circuits (NPCs), a new class of concept bottleneck models, comprise an attribute recognition model and a probabilistic circuit for reasoning. By integrating the outputs from these two modules, NPCs produce compositional and interpretable predictions. While offering enhanced interpretability and high performance on downstream tasks, the neural-network-based attribute recognition model remains a black box. This vulnerability allows adversarial attacks to manipulate attribute predictions by introducing carefully crafted, subtle perturbations to input images, potentially compromising the final predictions. In this paper, we theoretically analyze the adversarial robustness of NPC and demonstrate that it only depends on the robustness of the attribute recognition model and is independent of the robustness of the probabilistic circuit. Moreover, we propose RNPC, the first robust neural probabilistic circuit against adversarial attacks on the recognition module. RNPC introduces a novel class-wise integration for inference, ensuring a robust combination of outputs from the two modules. Our theoretical analysis demonstrates that RNPC exhibits provably improved adversarial robustness compared to NPC. Empirical results on image classification tasks show that RNPC achieves superior adversarial robustness compared to existing concept bottleneck models while maintaining high accuracy on benign inputs. The code is available at https://github.com/uiuctml/RNPC.

1 Introduction

Deep Neural Networks (DNNs) exhibit superior performance across a range of downstream tasks. However, DNNs are often criticized for their lack of interpretability, making it hard to understand the decision-making process, especially when they are deployed in high-stakes domains, such as legal justice and healthcare [1]. Concept Bottleneck Models (CBMs) [2–6] are a class of models that attempt to improve model interpretability by incorporating human-understandable binary concepts (*e.g.*, white color) as an intermediate layer, followed by simple predictors such as linear models. This bottleneck enables model predictions to be interpreted using the predicted concepts due to the simplicity of the linear predictors on top of the concepts. While demonstrating improved interpretability, CBMs usually suffer from a performance drop compared to DNNs. Recently, a new class of concept bottleneck models, Neural Probabilistic Circuits (NPCs) [7, 8], has been introduced, which offers a promising balance between model interpretability and task performance. NPC consists of an attribute recognition model and a probabilistic circuit [9]. The attribute recognition model predicts various interpretable categorical attributes (*e.g.*, color) from an input image. The probabilistic circuit supports tractable joint, marginal, and conditional inference over these attributes and the class variable. By integrating the probability of each instantiation of attributes and the conditional probability of a specific class given that instantiation, NPC generates the prediction score for the class.

Despite enhanced transparency in the model architecture, the attribute recognition model within NPC, implemented using a neural network, remains a black box. This raises the threat of malicious attacks

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

targeting the attribute recognition model. Adversarial attacks [10–12], a typical type of attack against neural networks, attempt to manipulate model predictions by applying carefully crafted, imperceptible perturbations to the input images. Such attacks against the attribute recognition model can mislead attribute predictions, potentially compromising NPC's performance on downstream tasks.

In this paper, we theoretically analyze NPC's robustness against these attacks, understanding how the robustness of individual modules affects that of the overall model. Surprisingly, we show that the robustness of the overall model only depends on the robustness of the attribute recognition model, and including a probabilistic circuit does not impact the robustness of the overall model. This is in sharp contrast to the compositional nature of NPC's estimation error, as demonstrated in Chen et al. [7, Theorem 2]. This means that adversarial robustness can be achieved for free by using probabilistic circuits on top of intermediate concepts, rather than the linear predictors used in conventional CBMs.

To further improve the adversarial robustness of NPC, we propose the Robust Neural Probabilistic Circuit (RNPC), which adopts the same model architecture as NPC while introducing a novel class-wise integration approach for inference. Specifically, we first partition the attribute space by class, where each class corresponds to a set of high-probability attribute instantiations, and then define the neighborhood for each class to allow perturbations. Rather than focusing on individual attribute instantiations, RNPC integrates the probability over the neighborhood of each class and the conditional probability of a target (class) given the high probability region of that class.

Theoretically, we show that such class-wise integration enables RNPC to achieve improved adversarial robustness compared to NPC. We also perform an analysis of RNPC's performance on benign inputs. Similar to NPC, the estimation error of RNPC is compositional and bounded by a linear combination of errors from its individual modules. Moreover, we provide an explicit characterization to quantify the trade-off between RNPC's adversarial robustness and benign performance.

Empirical results on diverse image classification datasets demonstrate that RNPC outperforms existing concept bottleneck models in robustness against three types of adversarial attacks across various attack budgets while maintaining high accuracy on benign inputs. Additionally, we conduct extensive ablation studies, including analyzing the impact of the number of attacked attributes and examining the effect of spurious correlations among various attributes.

Our main contributions are threefold. **1)** We propose the first robust neural probabilistic circuit, named RNPC, against adversarial attacks on the attribute recognition model. In particular, RNPC introduces a novel class-wise integration approach for inference, ensuring a robust combination of outputs from different modules. **2)** Theoretically, we demonstrate that: a) The robustness of NPC and RNPC depends only on the robustness of the attribute recognition model, and introducing a probabilistic circuit on top of the attribute recognition model is free for robustness. b) RNPC is guaranteed to achieve higher robustness than NPC under certain conditions. c) RNPC maintains a compositional estimation error on benign inputs. d) There exists a trade-off between RNPC's adversarial robustness and benign performance. **3)** Empirically, we show that RNPC achieves superior robustness against diverse adversarial attacks compared to various concept bottleneck models, while maintaining high accuracy on benign inputs.

2 Preliminaries

2.1 Neural probabilistic circuits

A Neural Probabilistic Circuit (NPC) [7] consists of an attribute recognition model and a probabilistic circuit. Let $X \in \mathcal{X}$, $Y \in \mathcal{Y}$, $A_k \in \mathcal{A}_k$ denote the input variable, the class variable, and the k-th attribute variable, respectively, with their lowercase letters representing the corresponding instantiations. Consider K attributes, $A_1, \ldots, A_K$, or $A_{1:K}$ in short. The neural-network-based attribute recognition model takes an image x as input and outputs probability vectors for various attributes. The k-th probability vector is denoted as $(\mathbb{P}_{\theta_k}(A_k = a_k \mid X = x))_{a_k \in \mathcal{A}_k}$, where θ_k represents the model parameters related to the k-th attribute. The probabilistic circuit [9] learns the joint distribution of Y and $A_{1:K}$, while also supporting tractable conditional inference such as $\mathbb{P}_w(Y \mid A_{1:K})$, where w represents the circuit's parameters. Combining the outputs from both the attribute recognition model and the probabilistic circuit, NPC's prediction score for class y is interpretable, which is the sum of the probability of each instantiation of attributes, weighted by the

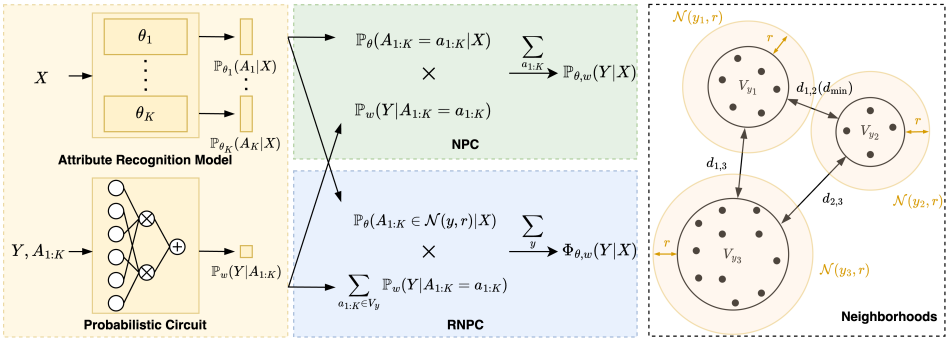

Figure 1: Left: Model architectures and inference procedures of NPC and RNPC. NPC and RNPC share an attribute recognition model and a probabilistic circuit. During inference, NPC employs a node-wise integration to integrate the outputs of the two modules, producing predictions for downstream tasks. In contrast, RNPC adopts a class-wise integration, which leads to more robust task predictions. **Right:** Illustration of three classes in the attribute space. V_y represents the set of attribute nodes with high probabilities. $d_{\min}$ is the minimum inter-class distance and the radius r is defined as $\lfloor \frac{d_{\min}-1}{2} \rfloor$. $\mathcal{N}(y, r)$ denotes the neighborhood of class y with radius r.

conditional probability of y given this instantiation, *i.e.*,

$$\mathbb{P}_{\theta,w}\left(Y = y \mid X = x\right) = \sum_{a_{1:K}} \prod_{k=1}^{K} \mathbb{P}_{\theta_k}\left(A_k = a_k \mid X = x\right) \cdot \mathbb{P}_w\left(Y = y \mid A_{1:K} = a_{1:K}\right), \tag{1}$$

where θ denotes all parameters of the attribute recognition model. An illustration of NPC is in Fig 1.

2.2 Threat model

Consider a white-box, norm-bounded, untargeted adversarial attack against the attribute recognition model. Given an input $(x, a_{1:K})$, the attacker seeks to find a perturbed input $\tilde{x}$ within an ℓ_p-norm ball of radius ℓ centered at x, such that the *attribute recognition model*'s predictions for *one or more* attributes become incorrect. Assume m attacked attributes $A_{i_1:i_m}$, the attack objective is $\max_{\tilde{x}\in\mathbb{B}_p(x,\ell)} \frac{1}{m}\sum_{k=1}^{m} \mathcal{L}\left(\mathbb{P}_{\theta_{i_k}}(A_{i_k} \mid X = \tilde{x}), a_{i_k}\right)$ where $\mathbb{B}_p(x,\ell) := \{\tilde{x} \in \mathcal{X} : \|\tilde{x} - x\|_p \leqslant \ell\}$ and $\mathcal{L}$ denotes a per-attribute loss function, which may vary depending on the chosen attack method.

3 Understanding the adversarial robustness of neural probabilistic circuits

Chen et al. [7] have shown that under Assumption 3.1 and 3.2, the *estimation error* of NPC is compositional, *i.e.*, it can be upper bounded by a linear combination of the error of the attribute recognition model and the error of the probabilistic circuit. In this section, we delve into the *adversarial robustness* of NPC, exploring how the adversarial robustness of the attribute recognition model as well as the probabilistic circuit affects that of NPC.

Assumption 3.1 (Sufficient attributes [7])**.** The class label Y and the input X are conditionally independent given the attributes $A_{1:K}$, *i.e.*, $Y \perp X \mid A_{1:K}$.

Assumption 3.2 (Complete information [7])**.** Given any input, all attributes are conditionally mutually independent, *i.e.*, $A_1 \perp A_2 \perp \cdots \perp A_K \mid X$.

Definition 3.3. The *prediction perturbation* of NPC against an adversarial attack on the attribute recognition model is defined as the worst-case total variance (TV) distance between the class distributions conditioned on the vanilla and perturbed inputs, *i.e.*,

$$\Delta^{\text{NPC}}_{\theta,w} := \mathbb{E}_X\left[\max_{\tilde{X}\in\mathbb{B}_p(X,\ell)} d_{\text{TV}}\left(\mathbb{P}_{\theta,w}(Y \mid X), \mathbb{P}_{\theta,w}(Y \mid \tilde{X})\right)\right].$$

The metric quantifies the adversarial robustness of NPC, with a lower value signifying stronger robustness. Based on this definition, the following theorem decomposes the adversarial robustness of NPC under the assumption of complete information.

[1]For uncluttered notation, we omit the dependency on ℓ if it is clear from the context.

Theorem 3.4 (Adversarial robustness of NPCs)**.** *Under Assumption 3.2, the prediction perturbation of NPC is bounded by the worst-case TV distance between the overall attribute distributions conditioned on the vanilla and perturbed inputs, which is further bounded by the sum of the worst-case TV distances for each attribute, i.e.,*

$$\Delta_{\theta,w}^{NPC} \leqslant \mathbb{E}_X \left[\max_{\tilde{X} \in \mathbb{B}_p(X,\ell)} d_{\mathrm{TV}} \left(\mathbb{P}_\theta \left(A_{1:K} \mid X \right), \mathbb{P}_\theta \left(A_{1:K} \mid \tilde{X} \right) \right) \right] \leqslant \sum_{k=1}^{K} \mathbb{E}_X \left[\max_{\tilde{X} \in \mathbb{B}_p(X,\ell)} d_{\mathrm{TV}} \left(\mathbb{P}_{\theta_k} \left(A_k \mid X \right), \mathbb{P}_{\theta_k} \left(A_k \mid \tilde{X} \right) \right) \right].$$

We denote the first bound as Λ_{NPC}. Theorem 3.4 demonstrates that the prediction perturbation of NPC is upper bounded by that of the attribute recognition model. Different from typical DNNs whose robustness is influenced by the robustness of each layer [13, 14], the robustness of NPC depends solely on that of the attribute recognition model. Adding a probabilistic circuit on top does not affect the robustness of NPC. Note that this is in sharp contrast to typical CBMs, where the linear-layer-based predictors adversely affect the robustness of the overall model [15].

4 Improving the adversarial robustness of neural probabilistic circuits

In this section, we propose Robust Neural Probabilistic Circuits (RNPCs). RNPCs introduce a novel approach for integrating the outputs from the attribute recognition model and the probabilistic circuit, leading to class predictions that are provably more robust than those of NPCs.

4.1 Notation and definitions

Let $D := \{(x, a_{1:K}, y)\}$ denote a dataset, and let $V := \{a_{1:K} : \mathbb{P}_D(A_{1:K} = a_{1:K}) \geqslant \gamma\}$ denote the corresponding set in the attribute space that has a high probability mass, specifically larger than a constant $\gamma \geq 0$. We partition V according to the most probable class $a_{1:K}$ is in, *i.e.*, if $y^* = \arg\max_{y \in \mathcal{Y}} \mathbb{P}_D(Y = y \mid A_{1:K} = a_{1:K})$, then $a_{1:K} \in V_{y^*}$. Overall, $V = \bigcup_{y \in \mathcal{Y}} V_y$ and $V_i \cap V_j = \emptyset,\ \forall i \neq j$. Let Ω denote the whole attribute space, and let $V^c := \Omega \backslash V$ denote the complement of V, which is the set of attribute instantiations with a probability mass at most γ. In the following, we mainly focus on the attribute set V and name each $a_{1:K}$ as an (attribute) node. The Hamming distance between two nodes, say $a_{1:K}$ and $a'_{1:K}$, is defined as the number of attributes in which the two nodes differ, *i.e.*, $\mathrm{Ham}(a_{1:K}, a'_{1:K}) := \sum_{k=1}^{K} \mathbb{I}(a_k \neq a'_k)$.

Definition 4.1. The *inter-class distance* between class i and class j is defined as the minimum Hamming distance between nodes of V_i and nodes of V_j, *i.e.*, $d_{i,j} := \min_{v_i \in V_i, v_j \in V_j} \{\mathrm{Ham}(v_i, v_j)\}$.

Definition 4.2. The *minimum inter-class distance* of an attribute set V is $d_{\min} := \min_{i,j \in \mathcal{Y}, i \neq j} \{d_{i,j}\}$. The *radius* of an attribute set V is $r := \lfloor \frac{d_{\min} - 1}{2} \rfloor$.

Definition 4.3. The *neighborhood* of class y with radius r is defined as the union of V_y and the nodes from V^c whose distance from V_y is not larger than r, *i.e.*, $\mathcal{N}(y, r) := V_y \bigcup \left\{ a^c_{1:K} \in V^c : \min_{a_{1:K} \in V_y} \mathrm{Ham}\left(a^c_{1:K}, a_{1:K}\right) \leqslant r \right\}$.

Figure 1 illustrates the high-probability attribute nodes and neighborhoods of three classes.

4.2 Robust neural probabilistic circuits

RNPCs employ the same model architecture as NPCs, but use a novel inference procedure that potentially leads to more robust predictions for downstream tasks.

Model architecture and training. The architecture of RNPC is the same as that of NPC, consisting of an attribute recognition model and a probabilistic circuit. These two modules are trained independently. Specifically, the attribute recognition model is trained by minimizing the sum of cross-entropy losses over all attributes, *i.e.*, $\min_\theta -\frac{1}{|D|} \sum_{(x, a_{1:K}) \in D} \sum_{k=1}^{K} \log \mathbb{P}_{\theta_k}(A_k = a_k \mid X = x)$. Following Chen et al. [7], the structure of the probabilistic circuit is learned using the LearnSPN algorithm [16], and its parameters are optimized with the CCCP algorithm [17]. Note that NPC and RNPC share the same trained attribute recognition model and the same learned probabilistic circuit; the only difference between them lies in the inference procedure.

Intuition. Interpreting $\prod_{k=1}^{K} \mathbb{P}_{\theta_k}\left(A_k = a_k \mid X\right)$ as the weight of an attribute node $a_{1:K}$ and $\mathbb{P}_w\left(Y \mid A_{1:K} = a_{1:K}\right)$ as the contribution of this node to Y, NPC's prediction $\mathbb{P}_{\theta,w}\left(Y \mid X\right)$ becomes

a weighted sum of all nodes' contributions to Y. While such predictions are generally accurate, they remain vulnerable to adversarial attacks. For example, consider an input image x depicting a no-entry sign, with the attribute label $a^*_{1:K} = (\text{red}, \text{circle}, \text{slash})$. For a well-trained model, both the weight of $a^*_{1:K}$ and its contribution to $y_{\text{no-entry}}$ are typically high, leading $y_{\text{no-entry}}$ to be the most probable class. However, if an attacker attacks any m attributes, say $A_{1:m}$, *i.e.*, perturbing the predicted probabilities for these attributes, the weight of $a^*_{1:K}$ will decrease due to the reduction in $\mathbb{P}_{\theta_k}(A_k = a^*_k \mid X)$ for $k \in [m]$. Consequently, the weight of the set $\{a_{1:K} : a_{1:m} \neq a^*_{1:m}, a_{m+1:K} = a^*_{m+1:K}\}$ increases, which lies within $\mathcal{N}(y_{\text{no-entry}}, r)$ when $m \leqslant r$. However, the contributions of these nodes (*e.g.*, $(\text{blue}, \text{circle}, \text{slash})$) to $y_{\text{no-entry}}$ could be very small, possibly causing $y_{\text{no-entry}}$ no longer the most probable class. Nevertheless, if those "shifted" weights can be aggregated and aligned with the high contribution of $a^*_{1:K}$ to $y_{\text{no-entry}}$, the adverse impact of the attacks can be alleviated.

Inference. Inspired by the example above, we propose a novel inference procedure that robustly integrates the output of the attribute recognition model and the output of the probabilistic circuit to produce the final predictions. In particular, instead of adopting NPC's node-wise integration, we introduce the following class-wise integration,

$$\Phi_{\theta,w}(Y \mid X) = \sum_{\tilde{y} \in \mathcal{Y}} \left(\mathbb{P}_\theta \left(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid X \right) \cdot \sum_{a_{1:K} \in V_{\tilde{y}}} \mathbb{P}_w \left(Y \mid A_{1:K} = a_{1:K} \right) \right). \tag{2}$$

Equation (2) characterizes an inference procedure that integrates the weight of (neighborhood of) each class with the contribution of (high-probability nodes of) this class to Y. Thus, RNPC's prediction $\Phi_{\theta,w}(Y \mid X)$ becomes a weighted sum of all classes' contributions to Y. In particular, $\Phi_{\theta,w}$ represents an unnormalized probability. The corresponding partition function and normalized probability are denoted as $Z_\theta(X) = \sum_{y \in \mathcal{Y}} \Phi_{\theta,w}(Y = y \mid X) = \sum_{\tilde{y} \in \mathcal{Y}} \left(\mathbb{P}_\theta \left(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid X \right) \cdot |V_{\tilde{y}}| \right)$ and $\hat{\Phi}_{\theta,w}(Y \mid X) = \Phi_{\theta,w}(Y \mid X) / Z_\theta(X)$, respectively.

In an adversarial attack, input perturbations result in perturbations in predicted attribute probabilities. If such an attack shifts the probabilities originally assigned to nodes of V_y to any other nodes within $\mathcal{N}(y, r)$, the class-wise integration ensures that the weight of class y is barely affected. Meanwhile, the conditional probabilities generated by the probabilistic circuit remain benign. Consequently, the predictions produced by RNPC are robust.

Complexity. Let $|f_k|$ denote the size of the k-th neural network in the attribute recognition model, and let $|S|$ be the size of the probabilistic circuit (*i.e.*, the number of edges). We have the following proposition that compares the computational complexities of inference in NPC and RNPC.

Proposition 4.4. *The computational complexities of inference in NPC and RNPC are respectively* $O\left(\sum_{k=1}^K |f_k| + |S| \cdot \prod_{k=1}^K |\mathcal{A}_k|\right)$ *and* $O\left(\sum_{k=1}^K |f_k| + |S| \cdot |V|\right)$, *with* $|V| \leqslant \prod_{k=1}^K |\mathcal{A}_k|$.

The detailed proof is deferred to Appendix E. Proposition 4.4 shows that RNPC is more efficient than NPC in terms of the inference complexity.

4.3 Theoretical analysis

In this section, we provide a theoretical analysis of the adversarial robustness and benign performance of RNPC. Similar to Definition 3.3, we first define a metric to quantify its adversarial robustness.

Definition 4.5. The *prediction perturbation* of RNPC against an adversarial attack on the attribute recognition model is defined as the worst-case TV distance between the class distributions conditioned on the vanilla and perturbed inputs, *i.e.*,

$$\Delta^{\text{RNPC}}_{\theta,w} := \mathbb{E}_X \left[\max_{\tilde{X} \in \mathbb{B}_p(X,\ell)} d_{\text{TV}} \left(\hat{\Phi}_{\theta,w}(Y \mid X), \hat{\Phi}_{\theta,w}(Y \mid \tilde{X}) \right) \right].$$

4.3.1 Adversarial robustness of RNPCs

Lemma 4.6 (Adversarial robustness of RNPCs)**.** *The prediction perturbation of RNPC is bounded by the worst-case change in probabilities within a neighborhood caused by the attack, i.e.,*

$$\Delta^{\textit{RNPC}}_{\theta,w} \leqslant \mathbb{E}_X \left[\max_{\tilde{X} \in \mathbb{B}_p(X,\ell)} \left\{ \max_{\tilde{y} \in \mathcal{Y}} \left| 1 - \frac{\mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid \tilde{X})}{\mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid X)} \right| \right\} \right].$$

We denote this bound as Λ_{RNPC}. Similar to NPC, the adversarial robustness of RNPC depends solely on the attribute recognition model and is not influenced by the probabilistic circuit. In particular, one can see from the above bound that, as long as the perturbation does not change the probabilities of attributes encoded in each ball $\mathcal{N}(\tilde{y}, r)$, the upper bound is 0, *i.e.*, the prediction is robust.

4.3.2 Comparison in adversarial robustness

NPC and RNPC adopt node-wise and class-wise integration approaches during inference, respectively, leading to different upper bounds on the prediction perturbation. Here, we investigate the relationship between these bounds and compare the adversarial robustness of NPC and RNPC.

Theorem 4.7 (Comparison in adversarial robustness)**.** *Consider a p-norm-bounded adversarial attack with a budget of ℓ. Assume the attribute recognition model f_θ is randomized and satisfies ϵ-Differential Privacy (DP) with respect to the p-norm. Let the probability of an attribute taking a specific value correspond to the expected model output, i.e., $\mathbb{P}_{\theta_k}(A_k = a_k \mid X) = \mathbb{E}[f_{\theta_k}(X)_{a_k}]$, where the expectation is taken over the randomness within the model. Under Assumption 3.2, the following holds: $\Lambda_{NPC} \leqslant \frac{|\mathcal{A}_1|\ldots|\mathcal{A}_K|}{2}\alpha_\epsilon$ and $\Lambda_{RNPC} \leqslant \alpha_\epsilon$, where $\alpha_\epsilon := \max\{1 - e^{-K\epsilon}, e^{K\epsilon} - 1\}$. Moreover, there exist instances where both inequalities simultaneously hold as equalities.*

The above theorem establishes the relationship between Λ_{NPC} and Λ_{RNPC} under the condition of DP. Specifically, compared to Λ_{RNPC}, Λ_{NPC} is bounded by an exponentially larger value that scales exponentially with the number of attributes. This larger bound potentially leads to significantly weaker adversarial robustness for NPC, highlighting the robustness improvement achieved by RNPC.

4.3.3 Benign task performance of RNPCs

In this section, we focus on RNPC's benign task performance and answer the following questions: *Does RNPC exhibit a compositional estimation error similar to NPC?* Furthermore, *while Theorem 4.7 indicates the robustness improvement of RNPC, is there a trade-off in its prediction accuracy?*

Proposition 4.8 (Optimal RNPCs)**.** *The optimal RNPC w.r.t. the expected TV distance between the predicted distribution $\hat{\Phi}_{\theta,w}(Y \mid X)$ and the ground-truth distribution $\mathbb{P}^*(Y \mid X)$ is $\hat{\Phi}^*(Y \mid X) := \Phi^*(Y \mid X)/Z^*(X)$, where*

$$\Phi^*(Y \mid X) := \sum_{\tilde{y}} \left(\mathbb{P}^*(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid X) \cdot \sum_{\tilde{a}_{1:K} \in V_{\tilde{y}}} \mathbb{P}^*(Y \mid A_{1:K} = \tilde{a}_{1:K}) \right),$$

and $Z^*(X)$ is the partition function. Here, $\mathbb{P}^*$ denotes the respective ground-truth distributions.

Definition 4.9. The *estimation error* of RNPC is defined as the expected TV distance between the predicted distribution and the optimal distribution, *i.e.*,

$$\hat{\varepsilon}^{\text{RNPC}}_{\theta,w} := \mathbb{E}_X \left[d_{\text{TV}} \left(\hat{\Phi}_{\theta,w}(Y \mid X), \hat{\Phi}^*(Y \mid X) \right) \right].$$

Theorem 4.10 (Compositional estimation error)**.** *The estimation error of RNPC is bounded by a linear combination of errors from the attribute recognition model and the probabilistic circuit, i.e.,*

$$\hat{\varepsilon}^{RNPC}_{\theta,w} \leqslant \mathbb{E}_X \left[\max_{\tilde{y}} \left| 1 - \frac{\mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid X)}{\mathbb{P}^*(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid X)} \right| \right] + \frac{2}{\gamma} d_{\text{TV}} \left(\mathbb{P}_w(Y, A_{1:K}), \mathbb{P}^*(Y, A_{1:K}) \right),$$

where $\mathbb{P}^*$ denotes respective ground-truth distributions.

Theorem 4.10 demonstrates that RNPC exhibits a compositional estimation error; improving either module *w.r.t.* its own error can enhance the benign performance of RNPC.

On the other hand, under Assumption 3.1 and 3.2, it is easy to show that the optimal NPC corresponds to the ground-truth distribution, which does not hold for RNPCs. Thus, we define the distance between the optimal RNPC and the ground-truth distribution as RNPCs' trade-off in benign performance.

Theorem 4.11 (Trade-off of RNPCs)**.** *The trade-off of RNPCs in benign performance, defined as the expected TV distance between the optimal RNPC $\hat{\Phi}^*(Y \mid X)$ and the ground-truth distribution $\mathbb{P}^*(Y \mid X)$, is bounded as follows,*

$$\mathbb{E}_X \left[d_{\text{TV}} \left(\hat{\Phi}^*(Y \mid X), \mathbb{P}^*(Y \mid X) \right) \right] \leqslant \mathbb{E}_X \left[\max_{\tilde{y}} \; d_{\text{TV}} \left(\bar{\mathbb{P}}^*(Y \mid A_{1:K} \in V_{\tilde{y}}), \mathbb{P}^*(Y \mid X) \right) \right],$$

where $\bar{\mathbb{P}}^*(Y \mid A_{1:K} \in V_{\tilde{y}}) := \frac{1}{|V_{\tilde{y}}|} \sum_{a_{1:K} \in V_{\tilde{y}}} \mathbb{P}^*(Y \mid A_{1:K} = a_{1:K})$ represents the average ground-truth conditional distribution of Y given $A_{1:K} \in V_{\tilde{y}}$.

Theorem 4.11 characterizes an upper bound on the trade-off, which is determined by the underlying data distributions and the partitioning of the attribute space, specifically $\{V_y\}$. Changes in these factors affect the optimal RNPC's distance from the ground-truth distribution, which can also be interpreted as the price of robustness paid by RNPCs.

5 Experiments

5.1 Experimental settings

Datasets. We create four image classification datasets. **1) MNIST-Add3:** This dataset is constructed from the MNIST dataset [18], following the standard processing procedures outlined in Manhaeve et al. [19], Bortolotti et al. [20]. Each image consists of a concatenation of three digit images, with each digit serving as one attribute. The task on this dataset is to predict the sum of these three digits. **2) MNIST-Add5:** This dataset is constructed similarly to MNIST-Add3, except that each image concatenates five digit images. **3) CelebA-Syn:** This dataset is synthesized based on the CelebA dataset [21] using StarGAN [22]. Each synthesized image demonstrates eight facial attributes, such as hair color. Each unique combination of attribute values is assigned a group number. The task on this dataset is to identify each image's group number. **4) GTSRB-Sub:** This dataset is a subset of the GTSRB dataset [23], where each image represents a traffic sign and is annotated with four attributes like color and shape. The task is to classify images into their corresponding sign types. To ensure certain minimum inter-class distances over the above datasets, we constrain the possible instantiations of attributes, *i.e.*, only images aligned with these instantiations are generated or sampled. For instance, in a three-dimensional attribute space, if the instantiations are limited to $\{(0,0,0),(2,2,2)\}$, the resulting minimum inter-class distance is 3. Using this approach, the minimum inter-class distances for MNIST-Add3, MNIST-Add5, CelebA-Syn, and GTSRB-Sub are set to 3, 5, 4, and 3.

Baseline models and architectures. We select three representative concept bottleneck models as baselines, including NPC [7], vanilla CBM [2], and DCR [3]. We adopt independent two-layer multilayer perceptrons (MLPs) to learn different attributes in both NPC and RNPC. To ensure a fair comparison, CBM[2] employs a two-layer MLP as its recognition module. DCR uses a similar architecture with the second layer replaced with its embedding layer.

Attack configurations. To validate the robustness of RNPC, we adopt attacks that can significantly compromise the model's predictions on the attacked attribute(s). Specifically, we use the ∞-norm-bounded PGD attack [11], the 2-norm-bounded PGD attack [11], and the 2-norm CW attack [12], all configured in the untargeted setting. Empirical results demonstrate that under our threat model, these attacks are sufficient to substantially reduce attribute prediction accuracy—often down to 0% under large norm bounds. **Evaluation metrics.** We adopt classification accuracy as the evaluation metric. Specifically, when testing on benign (adversarial) inputs, the accuracy is referred to as *benign (adversarial) accuracy*. Further details on the experimental settings can be found in Appendix F.

5.2 Main results

Benign accuracy. Table 1 shows that both RNPC and the baseline models perform exceptionally well across the four datasets, exhibiting benign accuracy approximating 100%. Notably, on these selected datasets, RNPC is comparable with NPC and even attains slightly higher benign accuracy on the MNIST-Add3 and MNIST-Add5 datasets. These empirical results indicate that RNPC's trade-off in benign accuracy could be negligible on these datasets.

Table 1: Benign accuracy (%) of CBM, DCR, NPC, and RNPC on four image classification datasets.

Dataset	CBM	DCR	NPC	RNPC
MNIST-Add3	99.02	98.54	99.32	**99.37**
MNIST-Add5	99.37	99.21	99.40	**99.51**
CelebA-Syn	99.83	99.45	**99.95**	**99.95**
GTSRB-Sub	99.42	99.42	**99.57**	99.49

[2]Unless otherwise specified, CBM refers to vanilla CBM.

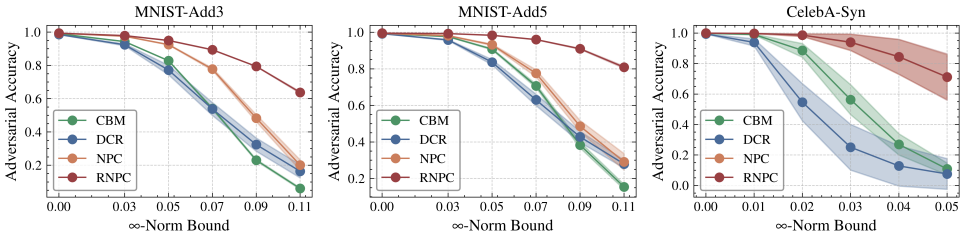

Figure 2: Adversarial accuracy of CBM, DCR, NPC, and RNPC under the ∞-norm-bounded PGD attack with varying norm bounds on the MNIST-Add3, MNIST-Add5, and CelebA-Syn datasets. The attacker attacks a single attribute at a time. The solid lines and the surrounding shaded regions represent the mean adversarial accuracy and the standard deviation, respectively, computed across all attacked attributes.

Adversarial accuracy. Figure 2 illustrates the adversarial accuracy of RNPC and the baseline models under the ∞-norm-bounded PGD attack with varying norm bounds. In this setup, the attacker attacks a single attribute at a time, generating an adversarial perturbation to distort the prediction result for that attribute. The solid lines and the surrounding shaded regions represent the mean adversarial accuracy and the standard deviation, respectively, computed across all attacked attributes. Additionally, Figure 6 in Appendix G.1 presents the accuracy of the attribute recognition model in predicting the attacked attribute, which drops to nearly 0% under large norm bounds. In Figure 2, we observe that across all datasets, the adversarial accuracy of all models decreases as the norm bound increases, demonstrating that stronger attacks cause greater harm to the models.

Importantly, NPC and RNPC consistently exhibit higher robustness compared to CBM and DCR, as their adversarial accuracy remains higher under attacks with any ∞-norm bound. This finding indicates that incorporating the probabilistic circuit into a model's architecture can strengthen its robustness, while the task predictors used in CBM and DCR might adversely impact model robustness.

Moreover, on the MNIST-Add3 and MNIST-Add5 datasets, RNPC significantly outperforms NPC, especially under attacks with larger norm bounds. For instance, on MNIST-Add5, when the ∞-norm bound reaches 0.11, NPC's adversarial accuracy drops below 40% whereas RNPC maintains an adversarial accuracy above 80%. These results demonstrate that RNPC provides superior robustness compared to NPC on these datasets, highlighting the effectiveness of the proposed class-wise integration approach. On the CelebA-Syn dataset, RNPC performs almost the same as NPC, with both showing high robustness even under attacks with large norm bounds. The performance against the 2-norm-bounded PGD and CW attacks is deferred to Appendix G.2.

5.3 Ablation studies

Impact of the number of attacked attributes. In Section 5.2, we analyze the models' performance under attacks targeting a single attribute. Here, we investigate the impact of the number of attacked attributes. To this end, we vary the number of attacked attributes for the ∞-norm-bounded PGD attack with a norm bound of 0.11. Specifically, on the MNIST-Add3 dataset, we attack the attributes "D1" ("D" stands for "Digit"), "D1, D2", and "D1, D2, D3", respectively. On MNIST-Add5, we additionally attack the attributes "D1, D2, D3, D4" and "D1, D2, D3, D4, D5". The adversarial accuracy under attacks with varying numbers of attacked attributes is shown in Figure 3 (a-b).

We observe that the adversarial accuracy of all models exhibits a downward trend as the number of attacked attributes increases, despite the perturbations remaining within the same norm bound. These results demonstrate that, compared to heavily perturbing the predicted probabilities of a single attribute, perturbing the predicted probabilities of multiple attributes—even if not as heavily—can have a more significant negative impact. Additionally, we find that RNPC consistently outperforms the baseline models by a margin under attacks across varying numbers of attacked attributes. These results underscore the high robustness of RNPC, even with a large number of attacked attributes.

Furthermore, we discover that, when the number of attacked attributes does not exceed the radius of a dataset, RNPC has a more distinct advantage over the baseline models. Specifically, RNPC achieves up to 45% higher adversarial accuracy than the best baseline model on MNIST-Add3 ($r = 1$) when *one* attribute is attacked, and 37% higher on MNIST-Add5 ($r = 2$) when *two* attributes are attacked. In contrast, when this number exceeds the radius, the advantage of RNPC tends to decrease.

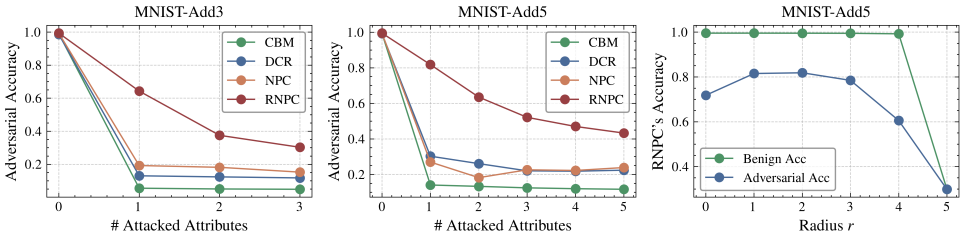

Figure 3: (a-b): Adversarial accuracy under the ∞-norm-bounded PGD attack (norm bound = 0.11) with varying numbers of attacked attributes on MNIST-Add3 and MNIST-Add5. **(c):** Performance of RNPC with different values of r on MNIST-Add5. This dataset has 5 attributes, and its attribute set has a radius of $r^* = 2$.

Impact of the radius. In Section 4.1, we define r as the radius of an attribute set, which is determined by the intrinsic structure of this set and fixed once the set is given. To avoid confusion, we denote this intrinsic radius of an attribute set as r^*. In Equation (2), r appears as a hyperparameter in the formulation of RNPC. Throughout the paper, we use $r = r^*$ by default. However, it can be adjusted, and in this section, we explore how varying r affects the performance of RNPC. We conduct this analysis on the MNIST-Add5 dataset, which has $K = 5$ attributes and an intrinsic radius $r^* = 2$. We vary RNPC's radius hyperparameter r in the range $[0, 5]$ and evaluate both benign and adversarial accuracy under an ∞-bounded PGD attack with a norm bound of 0.11.

We begin with benign accuracy. As r increases from 0 to 4, we observe a decreasing logit gap between the top-1 and top-2 predicted classes. Nevertheless, RNPC maintains near-perfect accuracy across this range (see Figure 3 (c)). However, when r reaches 5, accuracy drops sharply to 29.9%. This is expected because, at $r = 5$, the neighborhood $\mathcal{N}(\tilde{y}, r)$ spans the entire attribute space, *i.e.*, $\mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r)|X) = \mathbb{P}_\theta(A_{1:K} \in \Omega|X) = 1$. As a result, Equation (2) no longer incorporates any meaningful information from the input X.

Next, we examine adversarial performance. When $r \leqslant r^*$, decreasing r reduces adversarial accuracy. Conversely, increasing r beyond r^* also causes accuracy degradation. This indicates that the intrinsic radius r^* is optimal in the sense that deviating from r^*—in either direction—hurts robustness. Nevertheless, RNPC remains substantially more robust than other baseline models in these settings, as the baselines achieve less than 40% adversarial accuracy (see Figure 2 (middle)). Notably, when $r = 5$, adversarial accuracy falls to 29.9%, matching the benign accuracy and leading to a prediction perturbation of $\Delta^{\text{RNPC}}_{\theta,w} = 0$. This is consistent with our theoretical findings, as Lemma 4.6 indicates that the upper bound on perturbation becomes zero in this case. Overall, when $r \neq K$, RNPC demonstrates strong resilience to changes in r under benign settings; while its adversarial accuracy declines when r deviates from r^*, RNPC remains superior to other baseline models.

Impact of spurious correlations. Figure 4 (a) illustrates the adversarial accuracy of CBM, DCR, NPC, and RNPC on the GTSRB-Sub dataset, under the ∞-norm-bounded PGD attacks with varying norm bounds. We observe that RNPC and NPC exhibit similar performance, with both achieving higher adversarial accuracy compared to CBM and DCR under attacks with small norm bounds. However, as the norm bound increases, the advantage of RNPC and NPC gradually diminishes.

We hypothesize that, when the attribute recognition model is trained to recognize a specific attribute, the model might capture an unintended relationship between this attribute and the co-occurring features of other attributes. For instance, the shape "diamond" always co-occurs with the color "white" on the GTSRB-Sub dataset. Consequently, the model might rely on the features of the "diamond" shape to determine whether the color of an input image is "white". Such unintended relationships, known as *spurious correlations*, are a common phenomenon in neural networks [24, 25]. Due to the potential spurious correlations, attacking one attribute leads to attacking multiple attributes. We name such a phenomenon as *attack propagation*.

The results in Figure 4 (b) validate our hypothesis. Specifically, attacking any single attribute on the GTSRB-Sub dataset leads to a significant drop in the accuracy of recognizing other attributes. That is, although the attack targets only one attribute, the attack propagation induces more attacked attributes. As discussed in the study on the impact of the number of attacked attributes, when this number exceeds the radius of a dataset ($r = 1$ for GTSRB-Sub), the performance of RNPC could be compromised. Potential solutions for mitigating the attack propagation are discussed in Section 7. More ablation studies are deferred to Appendix G.3.

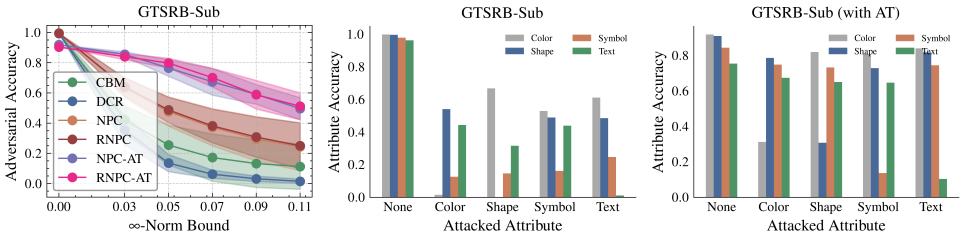

Figure 4: (a): Adversarial accuracy under the ∞-norm-bounded PGD attack with varying norm bounds on GTSRB-Sub. The attacker attacks a single attribute at a time. Solid lines and shaded regions represent the mean adversarial accuracy and the standard deviation, computed across all attacked attributes. **(b):** Accuracy of NPC and RNPC's attribute recognition model in recognizing various attributes under the ∞-norm-bounded PGD attack (norm bound = 0.11) targeting a different single attribute. **(c):** Same setting as (b), but using *adversarially trained* attribute recognition models for NPC and RNPC.

6 Related work

Robustness of concept bottleneck models. Exploring the robustness of CBMs against adversarial attacks is crucial for understanding their reliability in practical applications. Specifically, Sinha et al. [15] demonstrate that when the predicted concept probabilities are perturbed by attacks, CBMs often produce incorrect predictions. They further investigate *how to ensure that the predicted concept probabilities remain unchanged under adversarial attacks, thereby improving the robustness of CBMs*, and propose a training algorithm for CBMs based on adversarial training. Differently, our work admits the changes in the predicted concept probabilities and explores the question: *Can we make CBMs robust even when the predicted concept probabilities are perturbed by adversarial attacks?* A comprehensive literature review on CBMs, adversarial attacks, the robustness of CBMs, and the robustness of DNNs using probabilistic circuits is provided in Appendix A.

7 Discussion

In this section and Appendix B, we discuss the limitations of RNPC and explore potential solutions.

Mitigating the attack propagation effect. Section 5.3 shows that RNPC suffers from the attack propagation effect, which arises from spurious correlations learned by the attribute recognition model. Conventional solutions for mitigating these correlations include data augmentation [26, 27], counterfactual data generation [28, 29], etc. Here, we propose an adversarial-training-based approach that leverages adversarial examples to disentangle spurious correlations between attributes. Suppose the model relies on the feature of *white* to identify the shape as *diamond*. An adversarial example that perturbs the color attribute can shift its feature from *white* toward some another color, thereby weakening its association with *diamond*.

To achieve this, we generate adversarial examples that target a randomly selected attribute during training and train the model on both the adversarial and benign samples. As shown in Figure 4 (c), compared to models trained without adversarial training (see Figure 4 (b)), adversarially trained models exhibit significantly smaller drops in accuracy of other attributes when a specific attribute is attacked. This demonstrates a reduction in the attack propagation effect. Consequently, Figure 4 (a) shows that the robustness of both NPC and RNPC is enhanced, outperforming other baseline models across different attack norm bounds.

8 Conclusions

In this paper, we delve into the adversarial robustness of Neural Probabilistic Circuits (NPCs), showing that incorporating a probabilistic circuit into a model's architecture does not affect the robustness of the overall model. Moreover, we improve the robustness of NPCs by introducing a class-wise integration inference approach that produces robust predictions. Both theoretical and empirical results across various datasets and attacks demonstrate that the resulting model, named RNPC, achieves higher robustness. Due to the space limit, a more detailed conclusion is presented in Appendix C.

Acknowledgment

This work is supported by NSF IIS grant No. 2416897, NSF III grant No. 2504555 and an NSF CAREER Award No. 2442290. HZ would like to thank the support of a Google Research Scholar Award and Nvidia Academic Grant Award. The views and conclusions expressed in this paper are solely those of the authors and do not necessarily reflect the official policies or positions of the supporting companies and government agencies.

References

[1] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.*, 1(5):206–215, 2019.

[2] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 5338–5348, 2020.

[3] Pietro Barbiero, Gabriele Ciravegna, Francesco Giannini, Mateo Espinosa Zarlenga, Lucie Charlotte Magister, Alberto Tonda, Pietro Lio, Frédéric Precioso, Mateja Jamnik, and Giuseppe Marra. Interpretable neural-symbolic concept reasoning. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 1801–1825, 2023.

[4] Eunji Kim, Dahuin Jung, Sangha Park, Siwon Kim, and Sungroh Yoon. Probabilistic concept bottleneck models. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 16521–16540, 2023.

[5] Mert Yüksekgönül, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *ICLR*, 2023.

[6] Tuomas P. Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *ICLR*, 2023.

[7] Weixin Chen, Simon Yu, Huajie Shao, Lui Sha, and Han Zhao. Neural probabilistic circuits: Enabling compositional and interpretable predictions through logical reasoning. *arXiv preprint arXiv:2501.07021*, 2024.

[8] Weixin Chen, Simon Yu, Huajie Shao, Lui Sha, and Han Zhao. Neural probabilistic circuits: An overview. In *Eighth Workshop on Tractable Probabilistic Modeling*, 2025.

[9] Hoifung Poon and Pedro M. Domingos. Sum-product networks: A new deep architecture. In *UAI*, pages 337–346, 2011.

[10] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

[11] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

[12] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *IEEE SP*, pages 39–57, 2017.

[13] Xiaoyi Chen and Ni Zhang. Layer-wise adversarial training approach to improve adversarial robustness. In *IJCNN*, pages 1–8, 2020.

[14] Arash Rahnama, Andre T Nguyen, and Edward Raff. Robust design of deep neural networks against adversarial attacks based on lyapunov theory. In *CVPR*, pages 8178–8187, 2020.

[15] Sanchit Sinha, Mengdi Huai, Jianhui Sun, and Aidong Zhang. Understanding and enhancing robustness of concept-based models. In *AAAI*, pages 15127–15135, 2023.

[16] Robert Gens and Pedro M. Domingos. Learning the structure of sum-product networks. In *ICML*, pages 873–880, 2013.

[17] Han Zhao, Pascal Poupart, and Geoffrey J. Gordon. A unified approach for learning the parameters of sum-product networks. In *NeurIPS*, pages 433–441, 2016.

[18] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998.

[19] Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. Deepproblog: Neural probabilistic logic programming. In *NeurIPS*, pages 3753–3763, 2018.

[20] Samuele Bortolotti, Emanuele Marconato, Tommaso Carraro, Paolo Morettin, Emile van Krieken, Antonio Vergari, Stefano Teso, and Andrea Passerini. A neuro-symbolic benchmark suite for concept quality and reasoning shortcuts. In *NeurIPS (Track on Datasets and Benchmarks)*, 2024.

[21] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2015.

[22] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, pages 8789–8797, 2018.

[23] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: A multi-class classification competition. In *IJCNN*, pages 1453–1460, 2011.

[24] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nat. Mach. Intell.*, 2(11):665–673, 2020.

[25] Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, and Aidong Zhang. Spurious correlations in machine learning: A survey. *arXiv preprint arXiv:2402.12715*, 2024.

[26] Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *J. Big Data*, 6:60, 2019.

[27] Maximilian Ilse, Jakub M. Tomczak, and Patrick Forré. Selecting data augmentation for simulating interventions. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 4555–4562, 2021.

[28] Axel Sauer and Andreas Geiger. Counterfactual generative networks. In *ICLR*, 2021.

[29] Divyansh Kaushik, Eduard H. Hovy, and Zachary Chase Lipton. Learning the difference that makes A difference with counterfactually-augmented data. In *ICLR*, 2020.

[30] Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frédéric Precioso, Stefano Melacci, Adrian Weller, Pietro Lió, and Mateja Jamnik. Concept embedding models: Beyond the accuracy-explainability trade-off. In *NeurIPS*, 2022.

[31] Chih-Kuan Yeh, Been Kim, Sercan Ömer Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. In *NeurIPS*, 2020.

[32] Dmitry Kazhdan, Botty Dimanov, Mateja Jamnik, Pietro Liò, and Adrian Weller. Now you see me (CME): concept-based model extraction. In *CIKM*, volume 2699 of *CEUR Workshop Proceedings*, 2020.

[33] Anita Mahinpei, Justin Clark, Isaac Lage, Finale Doshi-Velez, and Weiwei Pan. Promises and pitfalls of black-box concept learning models. *arXiv preprint arXiv:2106.13314*, 2021.

[34] David M. Rodríguez, Manuel P. Cuéllar, and Diego Pedro Morales. Concept logic trees: enabling user interaction for transparent image classification and human-in-the-loop learning. *Appl. Intell.*, 54(5):3667–3679, 2024.

[35] Naveed Akhtar and Ajmal S. Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018.

[36] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *ICLR Workshop*, 2017.

[37] Nicolas Papernot, Patrick D. McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *EuroS&P*, pages 372–387, 2016.

[38] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *CVPR*, pages 2574–2582, 2016.

[39] Bader Rasheed, Mohamed Abdelhamid, Adil Khan, Igor Menezes, and Asad Masood Khattak. Exploring the impact of conceptual bottlenecks on adversarial robustness of deep neural networks. *IEEE Access*, 12:131323–131335, 2024.

[40] Nezihe Merve Gürel, Xiangyu Qi, Luka Rimanic, Ce Zhang, and Bo Li. Knowledge enhanced machine learning pipeline against diverse adversarial attacks. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 3976–3987, 2021.

[41] Zhuolin Yang, Zhikuan Zhao, Boxin Wang, Jiawei Zhang, Linyi Li, Hengzhi Pei, Bojan Karlas, Ji Liu, Heng Guo, Ce Zhang, and Bo Li. Improving certified robustness via statistical learning with logical reasoning. In *NeurIPS*, 2022.

[42] Jiawei Zhang, Linyi Li, Ce Zhang, and Bo Li. CARE: certifiably robust learning with reasoning via variational inference. In *SaTML*, pages 554–574, 2023.

[43] Mintong Kang, Nezihe Merve Gürel, Linyi Li, and Bo Li. COLEP: certifiably robust learning-reasoning conformal prediction via probabilistic circuits. In *ICLR*, 2024.

[44] Matthew Richardson and Pedro M. Domingos. Markov logic networks. *Mach. Learn.*, 62(1-2): 107–136, 2006.

[45] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[46] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763, 2021.

[48] Jingyi Xu, Zilu Zhang, Tal Friedman, Yitao Liang, and Guy Van den Broeck. A semantic loss function for deep learning with symbolic knowledge. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 5498–5507, 2018.

[49] Thomas Winters, Giuseppe Marra, Robin Manhaeve, and Luc De Raedt. Deepstochlog: Neural stochastic logic programming. In *AAAI*, pages 10090–10100, 2022.

[50] Emile van Krieken, Thiviyan Thanapalasingam, Jakub M. Tomczak, Frank van Harmelen, and Annette ten Teije. A-nesi: A scalable approximate method for probabilistic neurosymbolic inference. In *NeurIPS*, 2023.

[51] Jiani Huang, Ziyang Li, Binghong Chen, Karan Samel, Mayur Naik, Le Song, and Xujie Si. Scallop: From probabilistic deductive databases to scalable differentiable reasoning. In *NeurIPS*, pages 25134–25145, 2021.

[52] Kareem Ahmed, Stefano Teso, Kai-Wei Chang, Guy Van den Broeck, and Antonio Vergari. Semantic probabilistic layers for neuro-symbolic learning. In *NeurIPS*, 2022.

[53] Kareem Ahmed, Kai-Wei Chang, and Guy Van den Broeck. A pseudo-semantic loss for autoregressive models with logical constraints. In *NeurIPS*, 2023.

[54] Connor Pryor, Charles Dickens, Eriq Augustine, Alon Albalak, William Yang Wang, and Lise Getoor. Neupsl: Neural probabilistic soft logic. In *IJCAI*, pages 4145–4153, 2023.

[55] Zhun Yang, Adam Ishay, and Joohyung Lee. Neurasp: Embracing neural networks into answer set programming. In *IJCAI*, pages 1755–1762, 2020.

[56] Hoki Kim. Torchattacks: A pytorch repository for adversarial attacks. *arXiv preprint arXiv:2010.01950*, 2020.

[57] Mathias Lécuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *IEEE SP*, pages 656–672, 2019.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the abstract and introduction match theoretical and experimental results. Furthermore, claims in the introduction include the main contributions of the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix B comprehensively discusses the limitations of the work, including perspectives of scaling to concept-annotation-free datasets, augmenting the radius of a dataset, mitigating the attack propagation effect, and reducing the complete information assumption. Furthermore, this section provides potential solutions and highlights promising future directions.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Assumptions are stated clearly in each theorem or lemma. In addition, Appendix H provides a detailed proof for each theoretical result.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The contribution of this paper is primarily a new algorithm. Section 4.1 through 4.2 provide detailed steps for reproducing the algorithm. For replicating the experimental results, Section 5.1 in the main paper and Appendix F offer comprehensive information on the experimental setting, including dataset construction, implementation details such as model architecture and training parameters, attack configurations like norm bounds and step sizes, as well as evaluation metrics.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper provides the data and code in the supplementary materials, with sufficient instructions in the README file.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5.1 in the main paper and Appendix F offer comprehensive information on the experimental setting, including dataset construction, implementation details such as model architecture and training parameters, attack configurations like norm bounds and step sizes, as well as evaluation metrics.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: Results (*e.g.*, Figure 2, 3, 6, 7, 8) are accompanied by error bars. These bars are clearly explained in the text, in both figure captions and experimental analysis in particular.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Information on the computer resources (*e.g.*, type of compute workers, time of execution) is reported in Appendix E, in Table 2 in particular.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Broader impacts are discussed in Appendix D. We expect our work will not have any negative societal consequences if there is any.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The authors cite the original papers that produced the code package or dataset, and include the name of the license for each dataset.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper provides the new assets in the supplementary materials, with detailed documentation in the README file.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Appendix contents

A More related work

Concept bottleneck models. Concept bottleneck models (CBMs) were first introduced in Koh et al. [2]. These models are constructed by combining a concept recognition model with a task predictor. The concept recognition model, typically based on neural networks, takes an image as input and outputs probabilities for various high-level concepts, such as "red_color" and "white_color". The task predictor, often a linear model, determines the final class based on these predicted concept probabilities. Thanks to the simplicity of the linear predictor, the class predictions can be interpreted in terms of the predicted concept probabilities.

Follow-up works primarily focus on improving two aspects: the model's performance on downstream tasks and its interpretability. **1) Improving task performance:** Rather than using a layer of concept probabilities as the bottleneck, Zarlenga et al. [30], Yeh et al. [31], Kazhdan et al. [32], Kim et al. [4], Mahinpei et al. [33] propose the use of concept embeddings. Specifically, the concept recognition model generates high-dimensional embeddings for various concepts, and the task predictor determines the final class based on these embeddings. Since an embedding typically encodes more information than a single probability, these models often achieve better performance on downstream tasks compared to vanilla CBMs. However, the interpretability of these models is significantly compromised because the dimensions of the embeddings lack clear semantic meaning, and class predictions cannot be interpreted using the semantics of these embeddings. **2) Enhancing model interpretability:** To further enhance the interpretability of CBMs, several works introduce novel interpretable architectures for the task predictor. For instance, Barbiero et al. [3] propose a Deep Concept Reasoner (DCR) that allows class predictions to be interpreted through learned logical rules over the predicted concepts.

Similarly, Rodríguez et al. [34] employ a soft decision tree, where class predictions are generated by following specific branches within the tree. More recently, Chen et al. [7] explore the use of probabilistic circuits as task predictors, introducing a new model called Neural Probabilistic Circuits (NPCs). Specifically, NPC treats each group of concepts (*e.g.*, "red_color", "white_color") as one attribute (*e.g.*, "color") and consists of two modules: an attribute recognition model and a probabilistic circuit. The neural-network-based attribute recognition model takes an image as input and outputs probability vectors for various human-understandable attributes. The probabilistic circuit learns the joint distribution over the class variable and the attribute variables, while also supporting tractable marginal and conditional inference. Combining the outputs from the attribute recognition model and the probabilistic circuit, NPC produces class predictions that can be interpreted using the predicted probabilities of various attributes and the conditional dependencies between attributes and classes. Furthermore, Chen et al. [7] demonstrate that NPC exhibits a compositional estimation error, which is upper bounded by a linear combination of errors from its individual modules. Thanks to the integration of the probabilistic circuit, NPC achieves performance competitive with end-to-end DNNs while offering enhanced interpretability.

Adversarial attacks. Adversarial attacks [35] refer to the process of deliberately crafting small, often imperceptible, perturbations to input images with the aim of misleading neural networks into producing incorrect predictions. To ensure the imperceptibility of the crafted perturbations, adversarial attacks are typically norm-bounded, meaning the magnitude of the perturbation is constrained under a specified norm, such as L_1, L_2, or L_∞. Classical adversarial attacks include FGSM [10], BIM [36], PGD [11], CW [12], JSMA [37], and DeepFool [38]. The success of adversarial attacks underscores the vulnerabilities of neural networks, raising critical concerns about their robustness and reliability in practical applications.

According to the attacker's knowledge of the model, adversarial attacks can be categorized into *white-box* attacks and *black-box* attacks. In white-box attacks, the attacker has full access to the model's architecture and parameters, whereas black-box attacks only have access to the model's outputs, relying on query-based or transfer-based strategies to generate adversarial perturbations. According to the attacker's goal, adversarial attacks can be categorized into *targeted* attacks and *untargeted* attacks. In targeted attacks, the attacker aims to mislead the model into predicting a specific, incorrect label, while untargeted attacks focus on causing the model to produce any incorrect output. In this paper, we focus on white-box, untargeted adversarial attacks. In particular, we select the ∞-norm-bounded PGD attack [11], the 2-norm-bounded PGD attack [11], and the 2-norm-bounded CW attack [12].

Robustness of concept bottleneck models. By incorporating high-level, human-understandable concepts as an intermediate layer, Concept Bottleneck Models (CBMs) provide interpretable predictions that can be explained through the predicted concepts, thereby enhancing their reliability in practical applications. However, as the architectures of CBMs typically rely on neural networks, they can be vulnerable to adversarial attacks [10]. Exploring such vulnerabilities is crucial for understanding the potential threats underlying CBMs. In particular, Rasheed et al. [39], Sinha et al. [15], as well as our work, investigate the robustness of CBMs against adversarial attacks. Despite this shared focus, these studies address distinctly different problems.

Rasheed et al. [39] investigate the question: *Compared to DNNs, how robust are CBMs against adversarial attacks designed to mislead class predictions?* Their findings reveal that CBMs inherently exhibit higher robustness than their standard DNN counterparts. In contrast, Sinha et al. [15] and our work focus on adversarial attacks that target concept predictions rather than class predictions.

Sinha et al. [15] demonstrate that when the predicted concept probabilities are perturbed by adversarial attacks, CBMs often produce incorrect predictions. Given this vulnerability, they investigate *how to ensure that the predicted concept probabilities remain unchanged under adversarial attacks, thereby improving the robustness of CBMs*. To achieve this, they propose a training algorithm for CBMs based on adversarial training.

In contrast, our work admits the changes in the predicted concept probabilities and explores a different question: *Can we make a CBM, NPC in particular, robust even when the predicted concept probabilities are perturbed by adversarial attacks?* We demonstrate that by employing a class-wise integration approach, the final predictions of NPC are provably more robust. We also theoretically show that, unlike the linear model, the probabilistic circuit on top of the recognition module is free for robustness.

Improving robustness of end-to-end DNNs using probabilistic circuits. A line of research [40–43] explores leveraging probabilistic models—such as Markov logic networks [44] and probabilistic circuits [9]—to enhance the adversarial robustness of end-to-end DNNs. These approaches typically rely on a high-performance but vulnerable DNN to predict class labels for inputs that may contain adversarial perturbations. A probabilistic-model-based reasoning module is then used to correct potentially erroneous predictions made by the DNN. This predict-then-correct paradigm contrasts with our approach, which aims to build a robust and interpretable model from scratch, without relying on a high-performance DNN as the prime predictor.

This fundamental difference in objective also leads to differences in the problem setting. Following the framework of concept bottleneck models [2, 30], we assume access to a set of interpretable concepts/attributes that are sufficient to distinguish images from different classes. In contrast, the above research treats concept-based knowledge as auxiliary information used solely for correcting DNN predictions, and may consider only a limited set of attributes (*e.g.*, shape alone).

Among the research, Kang et al. [43] also employ probabilistic circuits. However, there are key differences in how probabilistic circuits are utilized. Specifically, our approach fully exploits the expressive power of probabilistic circuits: we learn smooth and decomposable circuits that represent the joint distribution over attributes and classes. This structure enables efficient inference—joint, marginal, and conditional distributions can all be computed with at most two forward passes through the circuit, highlighting the advantage of tractable probabilistic reasoning.

In contrast, Kang et al. [43] use probabilistic circuits less efficiently. Rather than modeling the joint distribution explicitly through the circuit's structure and edge weights, they define a factor function that computes the factor of each instantiation of attributes and class labels, which is essentially the exponential of the corresponding joint probability. These factors are treated as leaf nodes in the circuit. When a particular instantiation is provided as input, a product node connecting the corresponding factor leaf is activated, causing the circuit to output the associated joint probability. As a result, their circuit behaves more like an arithmetic circuit composed of sum and product nodes that performs arithmetic using factors, rather than a typical probabilistic circuit with embedded probabilistic semantics and tractable inference capabilities.

B More discussions

In this section, we discuss the limitations of RNPC and explore potential solutions and promising future directions.

Scaling to concept-annotation-free datasets. Consistent with standard concept bottleneck models [2, 3, 30], our work assumes access to concept annotations within the dataset. However, many image classification datasets such as CIFAR-10 [45] and ImageNet [46] lack such annotations, potentially limiting the applicability of our approach. Inspired by recent progress in label-free concept bottleneck models [5, 6], we can extend RNPC to annotation-free settings by using multi-modal models [47] to transfer concepts from other datasets or from natural language descriptions of concepts.

Scaling to more attributes. The computational complexity of RNPC is $O\left(\sum_{k=1}^{K}|f_k| + |S| \cdot |V|\right)$, where $|V| = \sum_{y\in\mathcal{Y}}|V_y| \in \left[|\mathcal{Y}|, \prod_{k=1}^{K}|\mathcal{A}_k|\right]$. This means the complexity can be as low as $|\mathcal{Y}|$ when each class only has one high-probability attribute assignment, but can grow exponentially with K in the worst case. Scalability remains a fundamental challenge for neuro-symbolic models, particularly those grounded in graphical model structures. Recent efforts have begun exploring approximation-based approaches to address this issue. That said, as demonstrated in Proposition 4.4, RNPC achieves a substantial reduction in computational complexity compared to NPC.

Augmenting the radius of a dataset. The inference procedure of RNPC relies on neighborhoods of various classes defined by a specific radius. Experimental results in Section 5.3 demonstrate that RNPC's performance under attacks is highly correlated with this radius. Specifically, when the number of attacked attributes does not exceed the radius, RNPC typically achieves high adversarial accuracy. Moreover, a larger radius generally leads to higher robustness against attacks. The radius is determined by the intrinsic properties of the dataset. For real-world datasets, the radius can be very small. For example, the original GTSRB dataset [23] has a minimum inter-class distance of 1,

resulting in a radius of 0. Such a small radius can limit RNPC's robustness against attacks. Conversely, augmenting this radius can enhance RNPC's performance.

A naive approach to augmenting the radius is to repeat certain attributes in the attribute annotations. For instance, the set $\{(0,0),(0,1)\}$ has a minimum inter-class distance of 1. If the second attribute is repeated, resulting in the set $\{(0,0,0),(0,1,1)\}$, the minimum inter-class distance increases by 1. More generally, if an attribute is repeated m times, the minimum inter-class distance $d_{\min}$ can increase by up to m. Another solution is to introduce new attributes that provide additional semantics for the downstream recognition tasks. These new attributes can potentially increase the radius. However, introducing new attributes requires additional annotation efforts, which may be challenging for large-scale datasets.

Reducing the complete information assumption. Assumption 3.2 assumes that each input image contains complete information about all attributes, suggesting that the attributes are conditionally mutually independent given the input. This assumption, commonly referred to as the conditional independence assumption, is widely adopted in neuro-symbolic learning frameworks [19, 48–51]. From a computational perspective, this assumption enables a factorization of the joint distribution over attributes into independent marginal distributions, *i.e.*, $\mathbb{P}_\theta(A_{1:K} \mid X) = \prod_{k=1}^{K} \mathbb{P}_{\theta_k}(A_k \mid X)$, which significantly reduces parameter complexity. Although this assumption is relatively mild, we consider the possibility of relaxing it in future work. When the complete information assumption does not hold, one could instead model the full joint distribution $\mathbb{P}_\theta(A_{1:K} \mid X)$ using a single expressive model [52–55]. While this approach introduces higher parameter complexity, it allows the model to capture interdependencies among attributes and accurately represent more complex joint patterns.

C Detailed conclusions

In this paper, we provide an understanding of the adversarial robustness of the Neural Probabilistic Circuit (NPC). Moreover, we improve the robustness of NPC by introducing a class-wise integration inference approach that produces robust predictions, and name the resulting model as RNPC.

Theorem 3.4 and Theorem 4.6 demonstrate the adversarial robustness of NPC and RNPC, respectively, showing that their robustness is upper bounded by the robustness of their attribute recognition models. These results also suggest that using a probabilistic circuit as the task predictor does not impact the robustness of the overall model. Theorem 4.7 compares these two bounds, showing that RNPC enhances the robustness of NPC under certain conditions. Furthermore, we analyze RNPC's benign performance on downstream tasks. Theorem 4.10 demonstrates the compositional estimation error of RNPC, showing that its estimation error is upper bounded by a linear combination of errors from its individual modules. Finally, Theorem 4.11 presents the distance between the optimal RNPC and the ground-truth distribution, revealing a trade-off between adversarial robustness and benign performance.

Empirical evaluations on diverse image classification datasets, under three types of adversarial attacks with varying norm bounds, demonstrate that RNPC achieves superior robustness compared with three baseline models while maintaining high accuracy on benign inputs. Additionally, ablation studies on the impact of the number of attacked attributes show that RNPC exhibits high robustness across varying numbers of attacked attributes. Besides, we observe that the spurious correlations captured by the attribute recognition model can induce attack propagation, which may compromise the robustness of RNPC.

Overall, with the proposed class-wise integration inference approach, RNPC achieves high robustness, capable of making correct class predictions even if the predicted attribute distributions are perturbed by adversarial attacks.

D Broader impact

This paper aims to understand and improve the robustness of neural probabilistic circuits against adversarial perturbations. Specifically, we propose a class-wise integration inference method and demonstrate that the resulting model is more robust under certain assumptions. Improving adversarial robustness enhances the trustworthiness of machine learning models, making them more reliable for deployment in real-world scenarios, especially in high-stakes applications. We expect that our method

Table 2: Comparison of training and inference time for NPC and RNPC across the MNIST-Add3, MNIST-Add5, and CelebA-Syn datasets.

Dataset	Phase	NPC	RNPC
MNIST-Add3	Training	78m	
	Inference	6.78s	4.78s
MNIST-Add5	Training	136m	
	Inference	16.52s	8.57s
CelebA-Syn	Training	381m	
	Inference	14.19s	8.58s

will inspire further research into enhancing the robustness of existing interpretable architectures or building an interpretable and robust model from scratch.

E Computational complexity comparison

This section presents a comparison of the computational complexity between NPC and RNPC in both the training and inference phases.

Training. Since NPC and RNPC share the same trained attribute recognition model and the same learned probabilistic circuit, their training complexities are identical. The practical training time across various datasets is reported in Table 2, with all experiments conducted using eight NVIDIA RTX A6000 GPUs.

Inference. Let $|f_k|$ denote the size of the k-th neural network in the attribute recognition model, and let $|S|$ denote the size of the probabilistic circuit (*i.e.*, the number of edges). During inference, given an input sample, a forward pass through all K neural networks in the attribute recognition model incurs a computational cost of $O(\sum_{k=1}^{K}|f_k|)$. According to the node-wise integration defined in Equation (1), NPC requires performing conditional inference over the probabilistic circuit $\prod_{k=1}^{K}|\mathcal{A}_k|$ times, resulting in the overall inference complexity:

$$O\left(\sum_{k=1}^{K}|f_k| + |S| \cdot \prod_{k=1}^{K}|\mathcal{A}_k|\right).$$

In contrast, according to the class-wise integration defined in Equation (2), RNPC only requires performing conditional inference over the probabilistic circuit $|V|$ times, where $V := \bigcup_{\tilde{y}\in\mathcal{Y}} V_{\tilde{y}}$, resulting in the overall inference complexity:

$$O\left(\sum_{k=1}^{K}|f_k| + |S| \cdot |V|\right).$$

By construction, $V \subseteq \Omega := \{a_{1:K}\}$, and therefore, $|V| \leqslant \prod_{k=1}^{K}|\mathcal{A}_k|$. It follows that $O(\sum_{k=1}^{K}|f_k| + |S| \cdot |V|) \leqslant O(\sum_{k=1}^{K}|f_k| + |S| \cdot \prod_{k=1}^{K}|\mathcal{A}_k|)$. Thus, RNPC is more efficient than NPC in terms of the inference complexity.

As expected, the practical inference time of RNPC, shown in Table 2, is consistently faster than NPC. All inference was performed on a single NVIDIA RTX A6000 GPU.

In summary, RNPC shares the same training complexity as NPC but offers better efficiency during inference.

F More details on experimental settings

F.1 Dataset construction

In Section 5.1, we describe the main properties (*e.g.*, attributes, downstream tasks) of various datasets. Here, we provide additional details about the construction process of each dataset.

MNIST-Add3 dataset. In this dataset, each image concatenates three digit images from the MNIST dataset [18] under the CC BY-SA 3.0 license. These digit images are applied with different transformations to introduce the domain shifts between them, which include rotations and color modifications. The three digits serve as the attributes for this dataset. The downstream task is to predict the sum of these attributes. To construct the dataset, we first identify an attribute set within the three-dimensional attribute space, ensuring a minimum inter-class distance of 3. Correspondingly, the radius of this attribute set is 1. The randomly selected attribute set V is {[6 3 7], [9 6 8], [0 2 4], [3 0 5], [5 5 1], [7 4 3], [2 7 6], [4 1 2], [1 9 0], [8 8 9]}. Each attribute node in V results in a unique attribute sum, leading to a total of 10 classes for the downstream task. Next, we generate images for each class by concatenating images of specific digits corresponding to the attributes. For example, to generate an image belonging to class 16, we concatenate a digit-6 image, a digit-3 image, and a digit-7 image in sequence. To introduce variability, we incorporate 1% labeling noise into the dataset. Specifically, this involves randomizing either the class labels or the attribute labels. In total, we generate 63,130 images and split them into training, validation, and testing sets by a ratio of 8:1:1. Example testing images are illustrated in Figure 5 (a).

MNIST-Add5 dataset. In this dataset, each image concatenates five digit images from the MNIST dataset [18] under the CC BY-SA 3.0 license. These digit images are applied with different transformations to introduce the domain shifts between them, which include rotations, color modifications, and blurring. The five digits serve as the attributes for this dataset. The downstream task is to predict the sum of these attributes. To construct the dataset, we first identify an attribute set within the five-dimensional attribute space, ensuring a minimum inter-class distance of 5. Correspondingly, the radius of this attribute set is 2. The randomly selected attribute set V is {[6 3 7 4 6], [0 9 5 3 1], [5 0 9 2 3], [2 7 8 5 4], [8 2 4 9 8], [7 5 0 6 0], [3 4 6 1 2], [4 1 3 0 5], [1 6 2 8 9], [9 8 1 7 7]}. These attribute nodes result in 7 different attribute sums, leading to a total of 7 classes. Note that a class (*e.g.*, class 26) may correspond to multiple attribute nodes in V. Next, we generate images for each class by concatenating images of specific digits corresponding to the attributes. For example, to generate an image belonging to class 32, we concatenate a digit-9 image, a digit-8 image, a digit-1 image, and two digit-7 images in sequence. To introduce variability, we incorporate 1% labeling noise into the dataset. Specifically, this involves randomizing either the class labels or the attribute labels. In total, we generate 63,130 images and split them into training, validation, and testing sets by a ratio of 8:1:1. Example testing images are illustrated in Figure 5 (b).

(a) MNIST-Add3 (b) MNIST-Add5

Figure 5: Examples of testing images from the MNIST-Add3 and MNIST-Add5 datasets. **(a)** A testing image from the MNIST-Add3 dataset, corresponding to the attribute node [6, 3, 7]. **(b)** A testing image from the MNIST-Add5 dataset, corresponding to the attribute node [5, 0, 9, 2, 3].

CelebA-Syn dataset. This dataset is constructed based on the CelebA dataset [21] under its non-commercial research license, which includes annotations for forty facial attributes. In particular, we select eight of them that are visually easy to distinguish, which are Color_Hair, Double_Chin, Eyeglasses, Heavy_Makeup, Mustache, Pale_Skin, Smiling, Young. Following a similar construction process as the above datasets, we first identify an attribute set within the eight-dimensional attribute space, ensuring a minimum inter-class distance of 4. Correspondingly, the radius of this attribute set is 1. The randomly selected attribute set V is {[2 1 0 0 0 1 0 0], [1 0 0 0 0 1 1 1], [1 1 1 1 0 1 1 0], [1 1 0 0 1 0 0 1], [2 0 1 1 1 1 0 1], [0 1 1 0 1 1 1 1], [0 0 0 1 0 0 0 1], [3 0 1 0 1 0 1 0], [0 0 0 1 1 1 1 0], [0 1 1 1 1 0 0 0]}[3]. Following Zarlenga et al. [30], each attribute node corresponds to a unique class, resulting in 10 classes in total. Next, we train a StarGAN [22] model and use it to synthesize images for each class. StarGAN is a powerful tool for transferring the attributes of input images to designated values. Specifically, taking a face image and a set of attribute values as input, the trained StarGAN generates a face image with attributes transferred to specified values. In total, we generate 50,000 training images, 10,000 validation images, and 9,990 testing images.

[3]For the attribute Color_Hair, values 0-3 correspond to Black_Hair, Blond_Hair, Brown_Hair, and Gray_Hair, respectively. For other attributes, *e.g.*, Double_Chin, 0 indicates the absence of the feature, while 1 indicates its presence.

GTSRB-Sub dataset. This dataset is derived from the GTSRB dataset [23] in accordance with its research-purpose terms, which contains images of German traffic signs along with class labels indicating the sign types. Chen et al. [7] annotate each image in GTSRB with four attributes: color, shape, symbol, and text. Each class corresponds to a distinct instantiation of attributes, *i.e.*, a unique attribute node. The attribute set corresponding to GTSRB has a minimum inter-class distance of 1, resulting in a radius of 0. To increase the minimum inter-class distance, we construct a subset of GTSRB by selecting and grouping specific classes. The final classes in this subset are: **1) Direction:** This class consists of the original classes—'regulatory–maximum-speed-limit-20', 'regulatory–maximum-speed-limit-30', 'regulatory–maximum-speed-limit-50', 'regulatory–maximum-speed-limit-60', 'regulatory–maximum-speed-limit-70', 'regulatory–maximum-speed-limit-80', 'regulatory–maximum-speed-limit-100', 'regulatory–maximum-speed-limit-120'. **2) Priority:** This class consists of the original class—'regulatory–priority-road'. **3) Speed:** This class consists of the original classes—'warning–other-danger', 'warning–double-curve-first-left', 'warning–uneven-road', 'warning–slippery-road-surface', 'warning–road-narrows-right', 'warning–roadworks', 'warning–traffic-signals', 'warning–pedestrians-crossing', 'warning–children', 'warning–bicycles-crossing', 'warning–ice-or-snow', 'warning–wild-animals'. **4) Warning:** This class consists of the original classes—'regulatory–turn-right-ahead', 'regulatory–turn-left-ahead', 'regulatory–go-straight', 'regulatory–go-straight-or-turn-right', 'regulatory–go-straight-or-turn-left', 'regulatory–keep-right', 'regulatory–keep-left', 'regulatory–roundabout'. Note that each class may correspond to multiple attribute nodes. This subset achieves a minimum inter-class distance of 3, with a corresponding radius of 1. In total, GTSRB-Sub contains 22,079 training images, 2,759 validation images, and 2,761 testing images.

F.2 Implementation details

NPC [7]. To strive for simplicity in experiments, we implement the attribute recognition model of NPC using a set of independent two-layer MLPs. Specifically, each MLP is used to identify one particular attribute. The attribute recognition model is trained using the sum of cross-entropy losses over all attributes. The training process is conducted with a batch size of 256 for 100 epochs, using the SGD optimizer. The probabilistic circuit of NPC is learned with the LearnSPN algorithm [16], with its parameters optimized via the CCCP algorithm [17]. Given the trained attribute recognition model and the learned probabilistic circuit, NPC integrates their outputs through node-wise integration to produce predictions for downstream tasks.

RNPC. RNPC uses the same trained attribute recognition model and learned probabilistic circuit as NPC. But different from NPC, RNPC adopts class-wise integration to produce predictions for downstream tasks.

CBM [2]. To ensure a fair comparison among the baselines, we implement the recognition module of CBM using a two-layer MLP. Following Koh et al. [2], the predictor module of CBM is implemented with a linear layer. CBM is trained using the weighted sum of the cross-entropy loss over concepts and the cross-entropy loss over classes. The training process is conducted with a batch size of 256 for 100 epochs, using the SGD optimizer.

DCR [3]. The recognition module of DCR is implemented with two layers: a linear layer followed by ReLU activation and an embedding layer defined in Zarlenga et al. [30]. The predictor module is implemented using the deep concept reasoner proposed in Barbiero et al. [3]. DCR is trained using the weighted sum of the cross-entropy loss over concepts and the cross-entropy loss over classes. The training process is conducted with a batch size of 256 for 100 epochs, using the SGD optimizer.

F.3 Attack configurations

The adversarial attacks used in this paper are implemented using the adversarial-attacks-pytorch library[4] [56].

∞-norm-bounded PGD attack [11]. For the MNIST-Add3, MNIST-Add5, and GTSRB-Sub datasets, the ∞-norm bounds are set to 0.03, 0.05, 0.07, 0.09, and 0.11. For the CelebA-Syn dataset, where the model demonstrates greater vulnerability to adversarial attacks, the ∞-norm bounds are set

[4]The library is available at https://github.com/Harry24k/adversarial-attacks-pytorch.

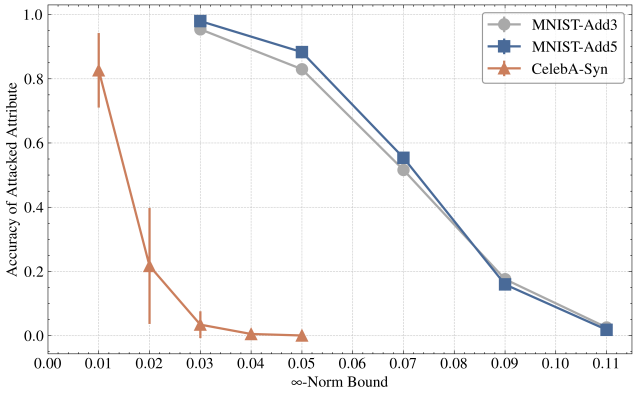

Figure 6: Accuracy of the attribute recognition model in predicting the attacked attribute under the ∞-norm-bounded PGD attack with varying norm bounds on the MNIST-Add3, MNIST-Add5, and CelebA-Syn datasets. The attacker attacks *a single attribute* at a time, generating an adversarial perturbation to distort the prediction result for that attribute. The curves show the mean accuracy of the attacked attribute across varying norm bounds, with error bars indicating the standard deviation computed over all attacked attributes.

to 0.01, 0.02, 0.03, 0.04, and 0.05. Across all datasets, we use a step size of 2/255 and perform 50 steps for the attack.

2-norm-bounded PGD attack [11]. For the MNIST-Add3 and MNIST-Add5 datasets, the 2-norm bounds are set to 3, 5, 7, 9, and 11. For the CelebA-Syn dataset, where the model demonstrates greater vulnerability to adversarial attacks, the 2-norm bounds are set to 1, 2, 3, 4, and 5. Across all datasets, we use a step size of 0.1*norm bound and perform 50 steps for the attack.

2-norm-bounded CW attack [12]. We employ the 2-norm-bounded CW attack with binary search. Across all datasets, we use a step size of 0.01 and perform 10 steps for the attack. The strength of the attack is varied by adjusting the number of binary search steps. Specifically, for the MNIST-Add3 and MNIST-Add5 datasets, the binary search steps are set to 3, 5, 7, 9, and 11. For the CelebA-Syn dataset, the steps are set to 1, 2, 3, 4, and 5. After the attack, we measure the 2-norm between the benign inputs and the perturbed inputs to quantify the magnitude of the perturbations.

G More experimental results

G.1 Performance for the attacked attribute

Consider a setting where the attacker attacks a single attribute at a time. Figure 6 illustrates the accuracy of the attribute recognition model in predicting the attacked attribute under the ∞-norm-bounded PGD attack with varying norm bounds on the MNIST-Add3, MNIST-Add5, and CelebA-Syn datasets. The curves show the mean accuracy of the attacked attribute across varying norm bounds, with error bars indicating the standard deviation computed over all attacked attributes.

As the norm bound increases, the accuracy for the attacked attribute consistently decreases, demonstrating that stronger adversarial perturbations more severely degrade the model's predictions. Notably, on all three datasets, the accuracy drops to nearly 0% at large norm bounds (*e.g.*, a bound of 0.11 for MNIST-Add3 and MNIST-Add5). This strong adversarial effect provides a compelling testbed for evaluating the robustness of RNPC.

G.2 Performance against more adversarial attacks

Performance against the 2-norm-bounded PGD attack. Figure 7 illustrates the adversarial accuracy of RNPC and the baseline models under the 2-norm-bounded PGD attack with varying norm bounds. According to these results, we can conduct a similar analysis and reach similar conclusions to those in Section 5.2. Specifically, we observe that on the MNIST-Add3, MNIST-Add5, and CelebA-Syn datasets, the adversarial accuracy of NPC and RNPC is consistently higher than that of CBM and DCR under attacks with any 2-norm bound. This finding indicates that including the

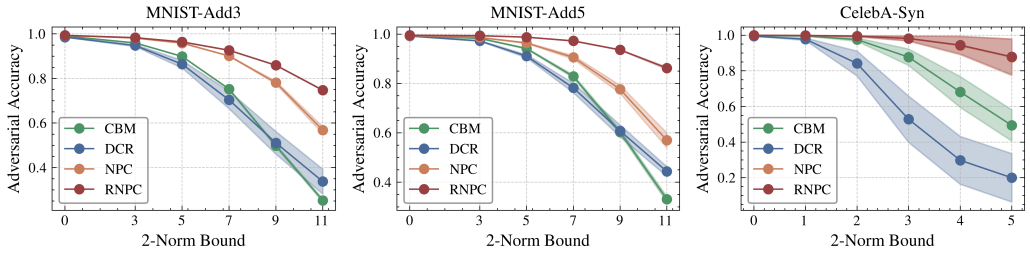

Figure 7: Adversarial accuracy of CBM, DCR, NPC, and RNPC under the 2-norm-bounded PGD attack with varying norm bounds on the MNIST-Add3, MNIST-Add5, and CelebA-Syn datasets.

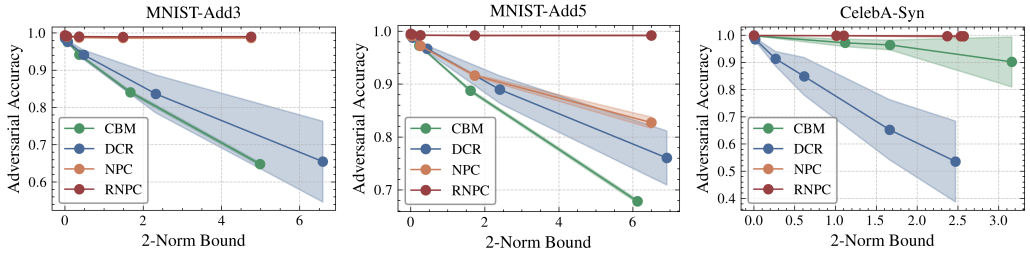

Figure 8: Adversarial accuracy of CBM, DCR, NPC, and RNPC under the 2-norm-bounded CW attack with varying norm bounds on the MNIST-Add3, MNIST-Add5, and CelebA-Syn datasets.

probabilistic circuit within a model's architecture potentially strengthens a model's robustness. In contrast, the task predictors used in CBM and DCR might compromise the model's robustness.

Furthermore, on the MNIST-Add3 and MNIST-Add5 datasets, RNPC outperforms NPC by a large margin, especially under attacks with larger norm bounds. For instance, on MNIST-Add5, when the 2-norm bound reaches 11, NPC's adversarial accuracy drops below 60% whereas RNPC maintains adversarial accuracy above 80%. These results demonstrate that RNPC provides superior robustness compared to NPC on these datasets, highlighting the effectiveness of the proposed class-wise integration approach. On the CelebA-Syn dataset, RNPC performs similarly to NPC, with both showing high robustness even under attacks with large norm bounds.

Performance against the 2-norm-bounded CW attack. Figure 8 illustrates the adversarial accuracy of RNPC and the baseline models under the 2-norm-bounded CW attack with varying norm bounds. We observe that NPC and RNPC perform similarly and robustly on the MNIST-Add3 and CelebA-Syn datasets, both reaching adversarial accuracy close to 100% under attacks with any 2-norm bound. In contrast, the adversarial accuracy of CBM and DCR decreases as the norm bound increases. This comparison demonstrates that NPC and RNPC are robust against the 2-norm-bounded CW attack, indicating the robustness enhancement enabled by the probabilistic circuit.

On the MNIST-Add5 dataset, however, NPC's adversarial accuracy also declines as the norm bound increases, while RNPC maintains adversarial accuracy close to 100%. These results demonstrate that RNPC is more robust than NPC, highlighting the robustness improvement achieved by the class-wise integration approach.

G.3 More ablation studies

Impact of Differential Privacy (DP). Theorem 4.7 indicates that the robustness of RNPC against a p-norm-bounded adversarial attack with a budget of ℓ is higher than that of NPC when the attribute recognition model satisfies ϵ-DP with respect to the p-norm. Here, we empirically validate whether this implication holds in practice.

To evaluate the robustness of NPC and RNPC, we conduct the 2-norm-bounded PGD attack (targeting a single attribute) with norm bounds of 3, 3, and 1 for the MNIST-Add3, MNIST-Add5, and CelebA-Syn datasets, respectively. Following Lécuyer et al. [57], DP within the attribute recognition model can be implemented by injecting noise after various layers. For simplicity, we directly add noise to the input images. Specifically, the noise is sampled from a Gaussian distribution with zero mean and

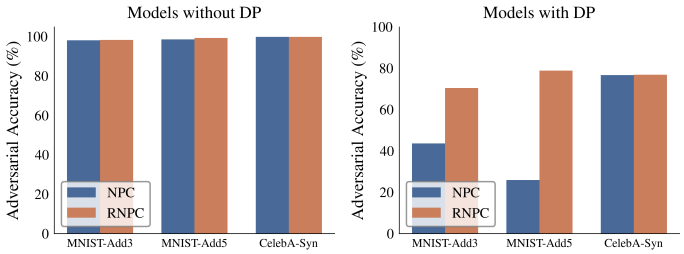

Figure 9: Adversarial accuracy of NPC and RNPC under the 2-norm-bounded PGD attack (targeting a single attribute) with norm bounds of 3, 3, and 1 for the MNIST-Add3, MNIST-Add5, and CelebA-Syn datasets, respectively. The left figure illustrates the performance of models without DP, *i.e.*, vanilla models, while the right figure illustrates the performance of models with their attribute recognition models satisfying DP.

standard deviation $\sigma = \sqrt{2\ln\left(\frac{1.25}{\delta}\right)\ell/\epsilon}$, where ℓ corresponds to the attack norm bound, ϵ is set to 0.5, and δ is chose as a small value (*e.g.*, 0.01). This ensures that the attribute recognition model approximately satisfies $(\epsilon, 0)$-DP. Additionally, we estimate $\mathbb{E}[f_{\theta_k}(X)_{a_k}]$ by computing the mean of $f_{\theta_k}(X)_{a_k}$ over 20 noise draws. The performance of NPC and RNPC under these conditions is illustrated in Figure 9.

Compared to models without DP (*i.e.*, vanilla models), the models satisfying DP generally exhibit lower adversarial accuracy across the three datasets due to the noise added to the inputs. Despite this, we observe that RNPC consistently outperforms NPC on the three datasets. Notably, on the MNIST-Add3 and MNIST-Add5 datasets, the adversarial accuracy of RNPC is higher than that of NPC by a large margin. These results demonstrate that while integrating DP into the attribute recognition model might compromise the overall performance on downstream tasks, it highlights the robustness enhancement achieved by RNPC, thereby validating the implication of Theorem 4.7.

H Theoretical results with omitted proofs

In this section, we provide more theoretical results and elaborate the proofs omitted in the main paper.

H.1 Adversarial robustness of NPCs

Theorem H.1 (Adversarial robustness of NPCs **(Restatement of Theorem 3.4)**). *Under Assumption 3.2, the prediction perturbation of NPC is bounded by the worst-case TV distance between the overall attribute distributions conditioned on the vanilla and perturbed inputs, which is further bounded by the sum of the worst-case TV distances for each attribute, i.e.,*

$$\Delta_{\theta,w}^{NPC} \leqslant \mathbb{E}_X\left[\max_{\tilde{X}\in\mathbb{B}_p(X,\ell)} d_{\mathrm{TV}}\left(\mathbb{P}_\theta\left(A_{1:K}\mid X\right), \mathbb{P}_\theta\left(A_{1:K}\mid \tilde{X}\right)\right)\right] \leqslant \sum_{k=1}^{K}\mathbb{E}_X\left[\max_{\tilde{X}\in\mathbb{B}_p(X,\ell)} d_{\mathrm{TV}}\left(\mathbb{P}_{\theta_k}\left(A_k\mid X\right), \mathbb{P}_{\theta_k}\left(A_k\mid \tilde{X}\right)\right)\right].$$

Proof. Under Assumption 3.2, $\mathbb{P}_\theta\left(A_{1:K}\mid X\right) = \prod_{k=1}^{K}\mathbb{P}_{\theta_k}(A_k\mid X)$. Therefore,

$$\begin{aligned}
&d_{\mathrm{TV}}\left(\mathbb{P}_{\theta,w}(Y\mid X), \mathbb{P}_{\theta,w}(Y\mid \tilde{X})\right)\\
&= \frac{1}{2}\sum_y \left|\mathbb{P}_{\theta,w}(Y=y\mid X) - \mathbb{P}_{\theta,w}(Y=y\mid \tilde{X})\right|\\
&\leqslant \frac{1}{2}\sum_y\sum_{a_{1:K}} \mathbb{P}_w(Y=y\mid A_{1:K}=a_{1:K})\cdot\left|\prod_{k=1}^{K}\mathbb{P}_{\theta_k}(A_k=a_k\mid X) - \prod_{k=1}^{K}\mathbb{P}_{\theta_k}(A_k=a_k\mid \tilde{X})\right|\\
&= \frac{1}{2}\sum_{a_{1:K}}\left|\prod_{k=1}^{K}\mathbb{P}_{\theta_k}(A_k=a_k\mid X) - \prod_{k=1}^{K}\mathbb{P}_{\theta_k}(A_k=a_k\mid \tilde{X})\right| = d_{\mathrm{TV}}\left(\mathbb{P}_\theta\left(A_{1:K}\mid X\right), \mathbb{P}_\theta\left(A_{1:K}\mid \tilde{X}\right)\right)\\
&\leqslant \frac{1}{2}\sum_{k=1}^{K}\sum_{a_k}\left|\mathbb{P}_{\theta_k}(A_k=a_k\mid X) - \mathbb{P}_{\theta_k}(A_k=a_k\mid \tilde{X})\right| = \sum_{k=1}^{K} d_{\mathrm{TV}}\left(\mathbb{P}_{\theta_k}(A_k\mid X), \mathbb{P}_{\theta_k}(A_k\mid \tilde{X})\right).
\end{aligned}$$

By successively applying the max and expectation operators to both sides, we complete the proof. □

H.2 Adversarial robustness of RNPCs

Lemma H.2 (Adversarial robustness of RNPCs **(Restatement of Lemma 4.6)**). *The prediction perturbation of RNPC is bounded by the worst-case change in probabilities within a neighborhood caused by the attack, i.e.,*

$$\Delta_{\theta,w}^{\mathit{RNPC}} \leqslant \mathbb{E}_X \left[\max_{\tilde{X} \in \mathbb{B}_p(X,\ell)} \left\{ \max_{\tilde{y} \in \mathcal{Y}} \left| 1 - \frac{\mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid \tilde{X})}{\mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid X)} \right| \right\} \right].$$

Proof. $\Delta_{\theta,w}^{\text{RNPC}}$ can be bounded as follows,

$$\begin{aligned}
&d_{\text{TV}}\left(\hat{\Phi}_{\theta,w}(Y \mid X), \hat{\Phi}_{\theta,w}(Y \mid \tilde{X})\right) \\
&= \frac{1}{2} \frac{1}{Z_\theta(X) \cdot Z_\theta(\tilde{X})} \sum_Y \left| Z_\theta(\tilde{X}) \cdot \Phi_{\theta,w}(Y \mid X) - Z_\theta(X) \cdot \Phi_{\theta,w}(Y \mid \tilde{X}) \right| \\
&= \frac{1}{2} \frac{1}{Z_\theta(X) \cdot Z_\theta(\tilde{X})} \sum_Y \left| Z_\theta(\tilde{X}) \cdot \Phi_{\theta,w}(Y \mid X) - Z_\theta(\tilde{X}) \cdot \Phi_{\theta,w}(Y \mid \tilde{X}) + Z_\theta(\tilde{X}) \cdot \Phi_{\theta,w}(Y \mid \tilde{X}) - Z_\theta(X) \cdot \Phi_{\theta,w}(Y \mid \tilde{X}) \right| \\
&\leqslant \frac{1}{2} \frac{1}{Z_\theta(X) \cdot Z_\theta(\tilde{X})} \sum_Y \left[Z_\theta(\tilde{X}) \cdot \left| \Phi_{\theta,w}(Y \mid X) - \Phi_{\theta,w}(Y \mid \tilde{X}) \right| + \Phi_{\theta,w}(Y \mid \tilde{X}) \cdot \left| Z_\theta(\tilde{X}) - Z_\theta(X) \right| \right] \\
&= \frac{1}{2} \frac{1}{Z_\theta(X)} \sum_Y \left| \Phi_{\theta,w}(Y \mid X) - \Phi_{\theta,w}(Y \mid \tilde{X}) \right| + \frac{1}{2} \frac{1}{Z_\theta(X) \cdot Z_\theta(\tilde{X})} \cdot Z_\theta(\tilde{X}) \cdot \left| Z_\theta(\tilde{X}) - Z_\theta(X) \right| \\
&= \frac{1}{2} \frac{1}{Z_\theta(X)} \left[\sum_Y \left| \Phi_{\theta,w}(Y \mid X) - \Phi_{\theta,w}(Y \mid \tilde{X}) \right| + \left| Z_\theta(\tilde{X}) - Z_\theta(X) \right| \right], \qquad (3)
\end{aligned}$$

where the penultimate equation is derived using $\sum_Y \Phi_{\theta,w}(Y \mid \tilde{X}) = Z_\theta(\tilde{X})$.

For the first interior term in Equation (3),

$$\left| \Phi_{\theta,w}(Y \mid X) - \Phi_{\theta,w}(Y \mid \tilde{X}) \right| \leqslant \sum_{\tilde{y}} \left| \mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid X) - \mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid \tilde{X}) \right| \cdot \sum_{a_{1:K} \in V_{\tilde{y}}} \mathbb{P}_w(Y \mid A_{1:K} = a_{1:K}),$$

$$\sum_Y \left| \Phi_{\theta,w}(Y \mid X) - \Phi_{\theta,w}(Y \mid \tilde{X}) \right| \leqslant \sum_{\tilde{y}} \left| \mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid X) - \mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid \tilde{X}) \right| \cdot |V_{\tilde{y}}|.$$

For the second interior term in Equation (3),

$$\left| Z_\theta(X) - Z_\theta(\tilde{X}) \right| \leqslant \sum_{\tilde{y}} \left| \mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid X) - \mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid \tilde{X}) \right| \cdot |V_{\tilde{y}}|.$$

By combining these two inequalities, Equation (3) is bounded by,

$$\begin{aligned}
\text{Equation(3)} &\leqslant \frac{1}{2} \frac{1}{Z_\theta(X)} \cdot 2 \sum_{\tilde{y}} \left| \mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid X) - \mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid \tilde{X}) \right| \cdot |V_{\tilde{y}}| \\
&= \frac{\sum_{\tilde{y}} \left| \mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid X) - \mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid \tilde{X}) \right| \cdot |V_{\tilde{y}}|}{\sum_{\tilde{y}} \mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid X) \cdot |V_{\tilde{y}}|} \\
&\leqslant \max_{\tilde{y}} \frac{\left| \mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid X) - \mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid \tilde{X}) \right|}{\mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid X)} \\
&= \max_{\tilde{y}} \left| 1 - \frac{\mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid \tilde{X})}{\mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid X)} \right|.
\end{aligned}$$

Consequently, $d_{\text{TV}}\left(\hat{\Phi}_{\theta,w}(Y \mid X), \hat{\Phi}_{\theta,w}(Y \mid \tilde{X})\right) \leqslant \max_{\tilde{y}} \left| 1 - \frac{\mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid \tilde{X})}{\mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid X)} \right|$. By successively applying the max and expectation operators to both sides, we complete the proof of Lemma 4.6. □

H.3 Comparison in adversarial robustness

Theorem H.3 (Comparison in adversarial robustness **(Restatement of Theorem 4.7)**). *Consider a p-norm-bounded adversarial attack with a budget of ℓ. Assume the attribute recognition model f_θ is randomized and satisfies ϵ-Differential Privacy (DP) with respect to the p-norm. Let the probability of an attribute taking a specific value correspond to the expected model output, i.e., $\mathbb{P}_{\theta_k}(A_k = a_k \mid X) = \mathbb{E}[f_{\theta_k}(X)_{a_k}]$, where the expectation is taken over the randomness within the*

model. Under Assumption 3.2, the following holds: $\Lambda_{NPC} \leqslant \frac{|\mathcal{A}_1|\ldots|\mathcal{A}_K|}{2}\alpha_\epsilon$ *and* $\Lambda_{RNPC} \leqslant \alpha_\epsilon$*, where* $\alpha_\epsilon := \max\{1 - e^{-K\epsilon}, e^{K\epsilon} - 1\}$*. Moreover, there exist instances where both inequalities hold as equalities.*

Proof. **Firstly**, we aim to prove that, under the given conditions, the following two statements hold for any $\tilde{X} \in \mathbb{B}_p(X, \ell)$, any $y \in \mathcal{Y}$, and any $a_{1:K} \in \mathcal{A}_1 \times \ldots \times \mathcal{A}_K$:

$$\left|\mathbb{P}_\theta(A_{1:K} = a_{1:K} \mid \tilde{X}) - \mathbb{P}_\theta(A_{1:K} = a_{1:K} \mid X)\right| \leqslant \left|\frac{\mathbb{P}_\theta(A_{1:K} = a_{1:K} \mid \tilde{X})}{\mathbb{P}_\theta(A_{1:K} = a_{1:K} \mid X)} - 1\right| \leqslant \alpha_\epsilon, \tag{4}$$

$$\left|\mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid \tilde{X}) - \mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid X)\right| \leqslant \left|\frac{\mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid \tilde{X})}{\mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid X)} - 1\right| \leqslant \alpha_\epsilon. \tag{5}$$

Given that the attribute recognition model satisfies ϵ-DP and using the expected output stability property of DP [57],

$$\mathbb{P}_{\theta_k}(A_k = a_k \mid X) \leqslant e^\epsilon \mathbb{P}_{\theta_k}(A_k = a_k \mid \tilde{X}).$$

Building on this, and under Assumption 3.2, the joint probability over all attributes $A_{1:K}$ is bounded by,

$$\mathbb{P}_\theta(A_{1:K} = a_{1:K} \mid X) = \prod_k \mathbb{P}_{\theta_k}(A_k = a_k \mid X) \leqslant (e^\epsilon)^K \prod_k \mathbb{P}_{\theta_k}(A_k = a_k \mid \tilde{X}) = e^{K\epsilon}\mathbb{P}_\theta(A_{1:K} = a_{1:K} \mid \tilde{X}).$$

Consequently, the probabilities within the neighborhood of any $\tilde{y} \in \mathcal{Y}$ are bounded by,

$$\mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid X) = \sum_{a_{1:K} \in \mathcal{N}(\tilde{y}, r)} \mathbb{P}_\theta(A_{1:K} = a_{1:K} \mid X) \leqslant \sum_{a_{1:K} \in \mathcal{N}(\tilde{y}, r)} e^{K\epsilon}\mathbb{P}_\theta(A_{1:K} = a_{1:K} \mid \tilde{X}) = e^{K\epsilon}\mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid \tilde{X}).$$

Therefore, $\frac{\mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid \tilde{X})}{\mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid X)} - 1 \geqslant e^{-K\epsilon} - 1$, and similarly, $\frac{\mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid \tilde{X})}{\mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid X)} - 1 \leqslant e^{K\epsilon} - 1$.

By combining these two inequalities, we obtain,

$$\left|\frac{\mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid \tilde{X})}{\mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid X)} - 1\right| \leqslant \alpha_\epsilon,$$

where $\alpha_\epsilon := \max\{1 - e^{-K\epsilon}, e^{K\epsilon} - 1\}$.

On the other hand, the following holds,

$$\left|\frac{\mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid \tilde{X})}{\mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid X)} - 1\right| = \frac{\left|\mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid \tilde{X}) - \mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid X)\right|}{\mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid X)} \geqslant \left|\mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid \tilde{X}) - \mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid X)\right|.$$

Hence, Equation (5) is satisfied. Following a similar approach, Equation (4) can also be proven to hold.

Secondly, we aim to prove that, when Equation (4) and Equation (5) are satisfied, $\Lambda_{\text{NPC}} \leqslant \frac{|\mathcal{A}_1|\ldots|\mathcal{A}_K|}{2}\alpha_\epsilon$ and $\Lambda_{\text{RNPC}} \leqslant \alpha_\epsilon$. In particular, the bound for Λ_{RNPC} is apparent based on Equation (5). Besides, for Λ_{NPC}, the following holds,

$$\Lambda_{\text{NPC}} = \mathbb{E}_X\left[\max_{\tilde{X} \in \mathbb{B}_p(X, \ell)} \left\{\frac{1}{2}\sum_{a_{1:K}} \left|\mathbb{P}_\theta(A_{1:K} = a_{1:K} \mid X) - \mathbb{P}_\theta(A_{1:K} = a_{1:K} \mid \tilde{X})\right|\right\}\right] \leqslant \frac{|\mathcal{A}_1| \ldots |\mathcal{A}_K|}{2}\alpha_\epsilon.$$

Therefore, we have proven that $\Lambda_{\text{NPC}} \leqslant \frac{|\mathcal{A}_1|\ldots|\mathcal{A}_K|}{2}\alpha_\epsilon$ and $\Lambda_{\text{RNPC}} \leqslant \alpha_\epsilon$ hold under the given conditions.

Finally, we aim to provide an instance showing that the bounds for Λ_{NPC} and Λ_{RNPC} can be simultaneously achieved.

Suppose a case where $\forall X \in \mathcal{X}$, there exists $\tilde{y} \in \mathcal{Y}$ such that $\mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid X) = 1$. There are $2n + 1$ nodes in $\mathcal{N}(\tilde{y}, r)$ and the probabilities of n of them are increased after attack, in particular, $\mathbb{P}_\theta(A_{1:K} = a_{1:K} \mid \tilde{X}) = \mathbb{P}_\theta(A_{1:K} = a_{1:K} \mid X) + \alpha_\epsilon$. In contrast, the remaining $n + 1$ of them are decreased, in particular, $\mathbb{P}_\theta(A_{1:K} = a_{1:K} \mid \tilde{X}) = \mathbb{P}_\theta(A_{1:K} = a_{1:K} \mid X) - \alpha_\epsilon$. Overall, $\mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid \tilde{X}) = \mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid X) - \alpha_\epsilon = 1 - \alpha_\epsilon$.

On the other hand, suppose there are $2m+1$ nodes in the complement set $\Omega\backslash\mathcal{N}(\tilde{y},r)$. The probabilities of m of them are decreased after attack, in particular, $\mathbb{P}_\theta(A_{1:K}=a_{1:K}\mid\tilde{X})=\mathbb{P}_\theta(A_{1:K}=a_{1:K}\mid X)-\alpha_\epsilon$. In contrast, the remaining $m+1$ of them are increased, in particular, $\mathbb{P}_\theta(A_{1:K}=a_{1:K}\mid\tilde{X})=\mathbb{P}_\theta(A_{1:K}=a_{1:K}\mid X)+\alpha_\epsilon$. Overall, $\mathbb{P}_\theta(A_{1:K}\in\Omega\backslash\mathcal{N}(\tilde{y},r)\mid\tilde{X})=\mathbb{P}_\theta(A_{1:K}\in\Omega\backslash\mathcal{N}(\tilde{y},r)\mid X)+\alpha_\epsilon=\alpha_\epsilon$.

In the above case, it is easy to show that $\Lambda_{\text{RNPC}}=\alpha_\epsilon$. In addition, we notice that $\forall a_{1:K},\ |\mathbb{P}_\theta(A_{1:K}=a_{1:K}\mid X)-\mathbb{P}_\theta(A_{1:K}=a_{1:K}\mid\tilde{X})|=\alpha_\epsilon$. Thus, $\Lambda_{\text{NPC}}=\frac{|\mathcal{A}_1|\ldots|\mathcal{A}_K|}{2}\alpha_\epsilon$. Therefore, in the case constructed above, the bounds for Λ_{NPC} and Λ_{RNPC} are simultaneously achieved. □

Theorem H.4 (Direct comparison in adversarial robustness)**.** *Assume that there exists $c\in(0,1)$ such that for all $X\in\mathcal{X}$ and $\tilde{y}\in\mathcal{Y}$, $\mathbb{P}_\theta(A_{1:K}\in\mathcal{N}(\tilde{y},r)\mid X)\geqslant c$. Then, the following inequality holds: $\Lambda_{RNPC}\leqslant\frac{1}{c}\Lambda_{NPC}$.*

Proof. By the definition of Λ_{RNPC} and the given conditions, the following holds,

$$\begin{aligned}\Lambda_{\text{RNPC}}&=\mathbb{E}_X\left[\max_{\tilde{y}}\frac{1}{\mathbb{P}_\theta(A_{1:K}\in\mathcal{N}(\tilde{y},r)\mid X)}\left|\mathbb{P}_\theta(A_{1:K}\in\mathcal{N}(\tilde{y},r)\mid X)-\mathbb{P}_\theta(A_{1:K}\in\mathcal{N}(\tilde{y},r)\mid\tilde{X})\right|\right]\\&\leqslant\frac{1}{c}\mathbb{E}_X\left[\max_{\tilde{y}}\left|\mathbb{P}_\theta(A_{1:K}\in\mathcal{N}(\tilde{y},r)\mid X)-\mathbb{P}_\theta(A_{1:K}\in\mathcal{N}(\tilde{y},r)\mid\tilde{X})\right|\right]\\&\leqslant\frac{1}{c}\mathbb{E}_X\left[d_{\text{TV}}\left(\mathbb{P}_\theta(A_{1:K}\mid X),\mathbb{P}_\theta(A_{1:K}\mid\tilde{X})\right)\right]=\frac{1}{c}\Lambda_{\text{NPC}}.\end{aligned}$$ □

Compared to Theorem 4.7, which compares the upper bounds for Λ_{NPC} and Λ_{RNPC}, Theorem H.4 provides a more direct relationship between them. Specifically, Theorem H.4 demonstrates that Λ_{RNPC} cannot exceed a fixed multiple of Λ_{NPC}, with the multiplier inversely proportional to the lower bound c of the neighborhood probabilities.

H.4 Benign task performance of RNPCs

Definition H.5. The *prediction error* of RNPC is defined as the expected TV distance between the predicted distribution and the ground-truth distribution, *i.e.*, $\varepsilon_{\theta,w}^{\text{RNPC}}:=\mathbb{E}_X\left[d_{\text{TV}}\left(\hat{\Phi}_{\theta,w}(Y\mid X),\mathbb{P}^*(Y\mid X)\right)\right]$.

Note that the definition of *prediction error* is different from that of *estimation error*. The latter is defined as the expected TV distance between the predicted distribution and the **optimal distribution**.

Theorem H.6 (Prediction error of RNPC)**.** *The prediction error of RNPC is bounded as follows,*

$$\begin{aligned}\varepsilon_{\theta,w}^{RNPC}\leqslant&\mathbb{E}_X\left[\min\left\{\max_{\tilde{y}}\left|\frac{\mathbb{P}_\theta(A_{1:K}\in\mathcal{N}(\tilde{y},r)\mid X)}{\mathbb{P}^*(A_{1:K}\in\mathcal{N}(\tilde{y},r)\mid X)}-1\right|,\max_{\tilde{y}}\left|\frac{\mathbb{P}^*(A_{1:K}\in\mathcal{N}(\tilde{y},r)\mid X)}{\mathbb{P}_\theta(A_{1:K}\in\mathcal{N}(\tilde{y},r)\mid X)}-1\right|\right\}\right]\\&+\frac{2}{\gamma}d_{\text{TV}}\left(\mathbb{P}_w(Y,A_{1:K}),\mathbb{P}^*(Y,A_{1:K})\right)+\mathbb{E}_X\left[\max_{\tilde{y}}d_{\text{TV}}\left(\bar{\mathbb{P}}^*(Y\mid A_{1:K}\in V_{\tilde{y}}),\mathbb{P}^*(Y\mid X)\right)\right],\end{aligned}$$

where $\bar{\mathbb{P}}^*(Y\mid A_{1:K}\in V_{\tilde{y}}):=\frac{1}{|V_{\tilde{y}}|}\sum_{a_{1:K}\in V_{\tilde{y}}}\mathbb{P}^*(Y\mid A_{1:K}=a_{1:K})$ represents the average ground-truth conditional distribution of Y given $A_{1:K}\in V_{\tilde{y}}$.

Proof. Define $\Phi^*(Y\mid X):=\sum_{\tilde{y}}\mathbb{P}^*(A_{1:K}\in\mathcal{N}(\tilde{y},r)\mid X)\cdot\sum_{a_{1:K}\in V_{\tilde{y}}}\mathbb{P}^*(Y\mid A_{1:K}=a_{1:K})$ and $Z^*(X):=\sum_{\tilde{y}}\mathbb{P}^*(A_{1:K}\in\mathcal{N}(\tilde{y},r)\mid X)\cdot|V_{\tilde{y}}|$.

By applying the triangle inequality, the following holds,

$$\varepsilon_{\theta,w}^{\text{RNPC}}\leqslant\mathbb{E}_X\left[d_{\text{TV}}\left(\frac{\Phi_{\theta,w}(Y\mid X)}{Z_\theta(X)},\frac{\Phi^*(Y\mid X)}{Z^*(X)}\right)\right]+\mathbb{E}_X\left[d_{\text{TV}}\left(\frac{\Phi^*(Y\mid X)}{Z^*(X)},\mathbb{P}^*(Y\mid X)\right)\right].\tag{6}$$

For the first term in Equation (6):

$$\begin{aligned}&d_{\text{TV}}\left(\frac{\Phi_{\theta,w}(Y\mid X)}{Z_\theta(X)},\frac{\Phi^*(Y\mid X)}{Z^*(X)}\right)=\mathbb{E}_X\left[\frac{1}{2}\sum_y\left|\frac{\Phi_{\theta,w}(Y=y\mid X)}{Z_\theta(X)}-\frac{\Phi^*(Y=y\mid X)}{Z^*(X)}\right|\right]\\&=\mathbb{E}_X\left[\frac{1}{2}\cdot\frac{1}{Z_\theta(X)\cdot Z^*(X)}\sum_y\left|Z^*(X)\cdot\Phi_{\theta,w}(Y=y\mid X)-Z_\theta(X)\cdot\Phi^*(Y=y\mid X)\right|\right].\end{aligned}$$

In particular, for the term $|Z^*(X)\cdot\Phi_{\theta,w}(Y=y\mid X)-Z_\theta(X)\cdot\Phi^*(Y=y\mid X)|$, we have,

$$
\begin{aligned}
&|Z^*(X)\cdot\Phi_{\theta,w}(Y=y\mid X)-Z_\theta(X)\cdot\Phi^*(Y=y\mid X)|\\
&=|Z^*(X)\cdot\Phi_{\theta,w}(Y=y\mid X)-Z_\theta(X)\cdot\Phi_{\theta,w}(Y=y\mid X)+Z_\theta(X)\cdot\Phi_{\theta,w}(Y=y\mid X)-Z_\theta(X)\cdot\Phi^*(Y=y\mid X)|\\
&\leqslant\Phi_{\theta,w}(Y=y\mid X)\cdot|Z_\theta(X)-Z^*(X)|+Z_\theta(X)\cdot|\Phi_{\theta,w}(Y=y\mid X)-\Phi^*(Y=y\mid X)|\,.
\end{aligned}
$$

Consequently, the following holds,

$$
\begin{aligned}
&d_{\mathrm{TV}}\left(\frac{\Phi_{\theta,w}(Y\mid X)}{Z_\theta(X)},\frac{\Phi^*(Y\mid X)}{Z^*(X)}\right)\\
&\leqslant\mathbb{E}_X\left[\frac{1}{2}\cdot\frac{1}{Z_\theta(X)\cdot Z^*(X)}\sum_y\Phi_{\theta,w}(Y=y\mid X)\cdot\left|Z_\theta(X)-Z^*(X)\right|+Z_\theta(X)\cdot\left|\Phi_{\theta,w}(Y=y\mid X)-\Phi^*(Y=y\mid X)\right|\right]\\
&=\mathbb{E}_X\left[\frac{1}{2}\cdot\frac{1}{Z^*(X)}|Z_\theta(X)-Z^*(X)|\right]+\mathbb{E}_X\left[\frac{1}{2}\cdot\frac{1}{Z^*(X)}\sum_y|\Phi_{\theta,w}(Y=y\mid X)-\Phi^*(Y=y\mid X)|\right].
\end{aligned}
$$

Moreover, we can bound the first term as follows,

$$
\begin{aligned}
\mathbb{E}_X\left[\frac{1}{2}\cdot\frac{1}{Z^*(X)}|Z_\theta(X)-Z^*(X)|\right]&\leqslant\mathbb{E}_X\left[\frac{1}{2}\cdot\frac{1}{Z^*(X)}\sum_{\tilde y}|\mathbb{P}_\theta(A_{1:K}\in\mathcal{N}(\tilde y,r)\mid X)-\mathbb{P}^*(A_{1:K}\in\mathcal{N}(\tilde y,r)\mid X)|\cdot|V_{\tilde y}|\right]\\
&\leqslant\mathbb{E}_X\left[\frac{1}{2}\max_{\tilde y}\left|\frac{\mathbb{P}_\theta(A_{1:K}\in\mathcal{N}(\tilde y,r)\mid X)}{\mathbb{P}^*(A_{1:K}\in\mathcal{N}(\tilde y,r)\mid X)}-1\right|\right],
\end{aligned}
$$

and bound the second term as follows,

$$
\begin{aligned}
&|\Phi_{\theta,w}(Y\mid X)-\Phi^*(Y\mid X)|\\
&=|\sum_{\tilde y}[\mathbb{P}_\theta(A_{1:K}\in\mathcal{N}(\tilde y,r)\mid X)\cdot\sum_{\tilde a_{1:K}\in V_{\tilde y}}\mathbb{P}_w(Y\mid A_{1:K}=\tilde a_{1:K})-\mathbb{P}^*(A_{1:K}\in\mathcal{N}(\tilde y,r)\mid X)\cdot\sum_{\tilde a_{1:K}\in V_{\tilde y}}\mathbb{P}_w(Y\mid A_{1:K}=\tilde a_{1:K})\\
&\quad+\mathbb{P}^*(A_{1:K}\in\mathcal{N}(\tilde y,r)\mid X)\cdot\sum_{\tilde a_{1:K}\in V_{\tilde y}}\mathbb{P}_w(Y\mid A_{1:K}=\tilde a_{1:K})-\mathbb{P}^*(A_{1:K}\in\mathcal{N}(\tilde y,r)\mid X)\cdot\sum_{\tilde a_{1:K}\in V_{\tilde y}}\mathbb{P}^*(Y\mid A_{1:K}=\tilde a_{1:K})]|\\
&\leqslant\sum_{\tilde y}|\mathbb{P}_\theta(A_{1:K}\in\mathcal{N}(\tilde y,r)\mid X)-\mathbb{P}^*(A_{1:K}\in\mathcal{N}(\tilde y,r)\mid X)|\cdot\sum_{\tilde a_{1:K}\in V_{\tilde y}}\mathbb{P}_w(Y\mid A_{1:K}=\tilde a_{1:K})\qquad(7)\\
&\quad+\sum_{\tilde y}\mathbb{P}^*(A_{1:K}\in\mathcal{N}(\tilde y,r)\mid X)\cdot|\sum_{\tilde a_{1:K}\in V_{\tilde y}}\mathbb{P}_w(Y\mid A_{1:K}=\tilde a_{1:K})-\sum_{\tilde a_{1:K}\in V_{\tilde y}}\mathbb{P}^*(Y\mid A_{1:K}=\tilde a_{1:K})|.\qquad(8)
\end{aligned}
$$

In particular, the following holds:

$$
\begin{aligned}
\mathbb{E}_X\left[\frac{1}{2}\cdot\frac{1}{Z^*(X)}\sum_y(Eq.(7))\right]&=\mathbb{E}_X\left[\frac{1}{2}\cdot\frac{1}{Z^*(X)}\sum_{\tilde y}|\mathbb{P}_\theta(A_{1:K}\in\mathcal{N}(\tilde y,r)\mid X)-\mathbb{P}^*(A_{1:K}\in\mathcal{N}(\tilde y,r)\mid X)|\cdot|V_{\tilde y}|\right]\\
&\leqslant\mathbb{E}_X\left[\frac{1}{2}\max_{\tilde y}\left|\frac{\mathbb{P}_\theta(A_{1:K}\in\mathcal{N}(\tilde y,r)\mid X)}{\mathbb{P}^*(A_{1:K}\in\mathcal{N}(\tilde y,r)\mid X)}-1\right|\right],
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{E}_X\left[\frac{1}{2}\cdot\frac{1}{Z^*(X)}\sum_y(Eq.(8))\right]\\
&\leqslant\mathbb{E}_X\left[\frac{1}{Z^*(X)}\sum_{\tilde y}\mathbb{P}^*(A_{1:K}\in\mathcal{N}(\tilde y,r)\mid X)\cdot\frac{1}{2}\sum_y\sum_{\tilde a_{1:K}\in V_{\tilde y}}|\mathbb{P}_w(Y=y\mid A_{1:K}=\tilde a_{1:K})-\mathbb{P}^*(Y=y\mid A_{1:K}=\tilde a_{1:K})|\right]\\
&=\mathbb{E}_X\left[\frac{1}{Z^*(X)}\sum_{\tilde y}\mathbb{P}^*(A_{1:K}\in\mathcal{N}(\tilde y,r)\mid X)\cdot\sum_{\tilde a_{1:K}\in V_{\tilde y}}d_{\mathrm{TV}}\left(\mathbb{P}_w(Y\mid A_{1:K}=\tilde a_{1:K}),\mathbb{P}^*(Y\mid A_{1:K}=\tilde a_{1:K})\right)\right]\\
&\leqslant\mathbb{E}_X\left[\frac{1}{Z^*(X)}\sum_{\tilde y}\mathbb{P}^*(A_{1:K}\in\mathcal{N}(\tilde y,r)\mid X)\cdot\sum_{\tilde a_{1:K}\in V_{\tilde y}}\frac{1}{\mathbb{P}^*(A_{1:K}=\tilde a_{1:K})}\cdot\sum_y|\mathbb{P}_w(Y=y,A_{1:K}=\tilde a_{1:K})-\mathbb{P}^*(Y=y,A_{1:K}=\tilde a_{1:K})|\right]\\
&\leqslant\frac{1}{\gamma}\max_{\tilde y}\frac{1}{|V_{\tilde y}|}\cdot\sum_{\tilde a_{1:K}\in V_{\tilde y}}\sum_y|\mathbb{P}_w(Y=y,A_{1:K}=\tilde a_{1:K})-\mathbb{P}^*(Y=y,A_{1:K}=\tilde a_{1:K})|\\
&\leqslant\frac{2}{\gamma}d_{\mathrm{TV}}\left(\mathbb{P}_w(Y,A_{1:K}),\mathbb{P}^*(Y,A_{1:K})\right).
\end{aligned}
$$

Within the above derivation, we utilize two facts. The first one is,

$$
\begin{aligned}
&d_{\mathrm{TV}}\left(\mathbb{P}_w(Y \mid A_{1:K}=a_{1:K}), \mathbb{P}^*(Y \mid A_{1:K}=a_{1:K})\right) \\
&= \frac{1}{2}\sum_y \left|\mathbb{P}_w(Y=y \mid A_{1:K}=a_{1:K}) - \mathbb{P}^*(Y=y \mid A_{1:K}=a_{1:K})\right| \\
&= \frac{1}{2}\cdot\frac{1}{\mathbb{P}^*(A_{1:K}=a_{1:K})}\sum_y \left|\mathbb{P}_w(Y=y \mid A_{1:K}=a_{1:K})\cdot\mathbb{P}^*(A_{1:K}=a_{1:K}) - \mathbb{P}_w(Y=y \mid A_{1:K}=a_{1:K})\cdot\mathbb{P}_w(A_{1:K}=a_{1:K})\right. \\
&\quad \left. + \mathbb{P}_w(Y=y \mid A_{1:K}=a_{1:K})\cdot\mathbb{P}_w(A_{1:K}=a_{1:K}) - \mathbb{P}^*(Y=y \mid A_{1:K}=a_{1:K})\cdot\mathbb{P}^*(A_{1:K}=a_{1:K})\right| \\
&\leqslant \frac{1}{2}\cdot\frac{1}{\mathbb{P}^*(A_{1:K}=a_{1:K})}\sum_y \mathbb{P}_w(Y=y \mid A_{1:K}=a_{1:K})\cdot\left|\mathbb{P}^*(A_{1:K}=a_{1:K}) - \mathbb{P}_w(A_{1:K}=a_{1:K})\right| \\
&\quad + \frac{1}{2}\cdot\frac{1}{\mathbb{P}^*(A_{1:K}=a_{1:K})}\sum_y \left|\mathbb{P}_w(Y=y, A_{1:K}=a_{1:K}) - \mathbb{P}^*(Y=y, A_{1:K}=a_{1:K})\right| \\
&\leqslant \frac{1}{\mathbb{P}^*(A_{1:K}=a_{1:K})}\sum_y \left|\mathbb{P}_w(Y=y, A_{1:K}=a_{1:K}) - \mathbb{P}^*(Y=y, A_{1:K}=a_{1:K})\right|.
\end{aligned}
$$

The second one is,

$$
\forall \tilde{a}_{1:K} \in V_{\tilde{y}}, \mathbb{P}^*(A_{1:K}=\tilde{a}_{1:K}) \geqslant \gamma.
$$

Combining the above, we have,

$$
d_{\mathrm{TV}}\left(\frac{\Phi_{\theta,w}(Y \mid X)}{Z_\theta(X)}, \frac{\Phi^*(Y \mid X)}{Z^*(X)}\right) \leqslant \mathbb{E}_X\left[\max_{\tilde{y}}\left|\frac{\mathbb{P}_\theta(A_{1:K}\in\mathcal{N}(\tilde{y},r) \mid X)}{\mathbb{P}^*(A_{1:K}\in\mathcal{N}(\tilde{y},r) \mid X)} - 1\right|\right] + \frac{2}{\gamma}d_{\mathrm{TV}}\left(\mathbb{P}_w(Y, A_{1:K}), \mathbb{P}^*(Y, A_{1:K})\right).
$$

Similarly, we can derive,

$$
d_{\mathrm{TV}}\left(\frac{\Phi_{\theta,w}(Y \mid X)}{Z_\theta(X)}, \frac{\Phi^*(Y \mid X)}{Z^*(X)}\right) \leqslant \mathbb{E}_X\left[\max_{\tilde{y}}\left|\frac{\mathbb{P}^*(A_{1:K}\in\mathcal{N}(\tilde{y},r) \mid X)}{\mathbb{P}_\theta(A_{1:K}\in\mathcal{N}(\tilde{y},r) \mid X)} - 1\right|\right] + \frac{2}{\gamma}d_{\mathrm{TV}}\left(\mathbb{P}_w(Y, A_{1:K}), \mathbb{P}^*(Y, A_{1:K})\right).
$$

Therefore, the following holds,

$$
\begin{aligned}
d_{\mathrm{TV}}\left(\frac{\Phi_{\theta,w}(Y \mid X)}{Z_\theta(X)}, \frac{\Phi^*(Y \mid X)}{Z^*(X)}\right) \leqslant \mathbb{E}_X\left[\min\left\{\max_{\tilde{y}}\left|\frac{\mathbb{P}^*(A_{1:K}\in\mathcal{N}(\tilde{y},r) \mid X)}{\mathbb{P}_\theta(A_{1:K}\in\mathcal{N}(\tilde{y},r) \mid X)} - 1\right|, \max_{\tilde{y}}\left|\frac{\mathbb{P}_\theta(A_{1:K}\in\mathcal{N}(\tilde{y},r) \mid X)}{\mathbb{P}^*(A_{1:K}\in\mathcal{N}(\tilde{y},r) \mid X)} - 1\right|\right\}\right] \\
+ \frac{2}{\gamma}d_{\mathrm{TV}}\left(\mathbb{P}_w(Y, A_{1:K}), \mathbb{P}^*(Y, A_{1:K})\right).
\end{aligned}
$$

For the second term in Equation (6): The following holds,

$$
\mathbb{E}_X\left[d_{\mathrm{TV}}\left(\frac{\Phi^*(Y \mid X)}{Z^*(X)}, \mathbb{P}^*(Y \mid X)\right)\right] \leqslant \mathbb{E}_X\left[\max_{\tilde{y}} d_{\mathrm{TV}}\left(\bar{\mathbb{P}}^*(Y \mid A_{1:K}\in V_{\tilde{y}}), \mathbb{P}^*(Y \mid X)\right)\right],
$$

because:

$$
\begin{aligned}
d_{\mathrm{TV}}\left(\frac{\Phi^*(Y \mid X)}{Z^*(X)}, \mathbb{P}^*(Y \mid X)\right) &= \frac{1}{2}\sum_Y \left|\frac{\sum_{\tilde{y}}\mathbb{P}^*(A_{1:K}\in\mathcal{N}(\tilde{y},r) \mid X)\cdot\sum_{a_{1:K}\in V_{\tilde{y}}}\mathbb{P}^*(Y \mid A_{1:K}=a_{1:K})}{\sum_{\tilde{y}}\mathbb{P}^*(A_{1:K}\in\mathcal{N}(\tilde{y},r) \mid X)\cdot|V_{\tilde{y}}|} - \mathbb{P}^*(Y \mid X)\right| \\
&= \frac{1}{2}\sum_Y \left|\frac{\sum_{\tilde{y}}\mathbb{P}^*(A_{1:K}\in\mathcal{N}(\tilde{y},r) \mid X)\cdot|V_{\tilde{y}}|\cdot\bar{\mathbb{P}}^*(Y \mid A_{1:K}\in V_{\tilde{y}})}{\sum_{\tilde{y}}\mathbb{P}^*(A_{1:K}\in\mathcal{N}(\tilde{y},r) \mid X)\cdot|V_{\tilde{y}}|} - \mathbb{P}^*(Y \mid X)\right|. \quad (9)
\end{aligned}
$$

Define $W_{\tilde{y}} := \mathbb{P}^*(A_{1:K}\in\mathcal{N}(\tilde{y},r) \mid X)\cdot|V_{\tilde{y}}|$ and $\alpha_{\tilde{y}} := \frac{W_{\tilde{y}}}{\sum_{y'}W_{y'}}$. We have $\sum_{\tilde{y}}\alpha_{\tilde{y}} = 1$.

Then, for Equation (9),

$$
\begin{aligned}
\text{Equation(9)} = \frac{1}{2}\sum_Y \left|\sum_{\tilde{y}}\alpha_{\tilde{y}}\bar{\mathbb{P}}^*(Y \mid A_{1:K}\in V_{\tilde{y}}) - \mathbb{P}^*(Y \mid X)\right| &\leqslant \frac{1}{2}\sum_Y\sum_{\tilde{y}}\alpha_{\tilde{y}}\left|\bar{\mathbb{P}}^*(Y \mid A_{1:K}\in V_{\tilde{y}}) - \mathbb{P}^*(Y \mid X)\right| \\
&= \sum_{\tilde{y}}\alpha_{\tilde{y}}\, d_{TV}\left(\bar{\mathbb{P}}^*(Y \mid A_{1:K}\in V_{\tilde{y}}), \mathbb{P}^*(Y \mid X)\right) \\
&\leqslant \max_{\tilde{y}} d_{TV}\left(\bar{\mathbb{P}}^*(Y \mid A_{1:K}\in V_{\tilde{y}}), \mathbb{P}^*(Y \mid X)\right).
\end{aligned}
$$

By combining the bounds for the first and second terms in Equation (6), we establish the bound for $\varepsilon^{\mathrm{RNPC}}_{\theta,w}$. □

Proposition H.7 (Optimal RNPCs (**Restatement of Proposition 4.8**)). *The optimal RNPC w.r.t. the prediction error $\varepsilon^{RNPC}_{\theta,w}$ is $\hat{\Phi}^*(Y \mid X) := \frac{\Phi^*(Y|X)}{Z^*(X)}$, where*

$$
\begin{aligned}
\Phi^*(Y \mid X) &:= \sum_{\tilde{y}}\mathbb{P}^*(A_{1:K}\in\mathcal{N}(\tilde{y},r) \mid X)\cdot\sum_{a_{1:K}\in V_{\tilde{y}}}\mathbb{P}^*(Y \mid A_{1:K}=a_{1:K}), \\
Z^*(X) &:= \sum_{\tilde{y}}\mathbb{P}^*(A_{1:K}\in\mathcal{N}(\tilde{y},r) \mid X)\cdot|V_{\tilde{y}}|.
\end{aligned}
$$

Proof. According to Theoremn H.6, the minimum of the bound for $\varepsilon_{\theta,w}^{\text{RNPC}}$ is achieved when $\mathbb{P}_\theta(A_{1:K} \mid X) = \mathbb{P}^*(A_{1:K} \mid X)$ and $\mathbb{P}_w(Y, A_{1:K}) = \mathbb{P}^*(Y, A_{1:K})$. In this case, $\Phi_{\theta,w}(Y \mid X) = \Phi^*(Y \mid X)$ and $Z_\theta(X) = Z^*(X)$.

□

Theorem H.8 (Trade-off of RNPCs **(Restatement of Theoremn 4.11)**). *The trade-off of RNPCs in benign performance, defined as the expected TV distance between the optimal RNPC $\hat{\Phi}^*(Y \mid X)$ and the ground-truth distribution $\mathbb{P}^*(Y \mid X)$, is bounded as follows,*

$$\mathbb{E}_X\left[d_{\text{TV}}\left(\hat{\Phi}^*(Y \mid X), \mathbb{P}^*(Y \mid X)\right)\right] \leqslant \mathbb{E}_X\left[\max_{\tilde{y}} d_{\text{TV}}\left(\bar{\mathbb{P}}^*(Y \mid A_{1:K} \in V_{\tilde{y}}), \mathbb{P}^*(Y \mid X)\right)\right],$$

where $\bar{\mathbb{P}}^*(Y \mid A_{1:K} \in V_{\tilde{y}}) := \frac{1}{|V_{\tilde{y}}|}\sum_{a_{1:K} \in V_{\tilde{y}}} \mathbb{P}^*(Y \mid A_{1:K} = a_{1:K})$ represents the average ground-truth conditional distribution of Y given $A_{1:K} \in V_{\tilde{y}}$.

Proof. See proof for Theorem 6. □

Theorem H.9 (Compositional estimation error **(Extension of Theorem 4.10)**). *The estimation error of RNPC is bounded by a linear combination of the errors from the attribute recognition model and the probabilistic circuit, i.e.,*

$$\begin{aligned}\hat{\varepsilon}_{\theta,w}^{\text{RNPC}} \leqslant \mathbb{E}_X\left[\min\left\{\max_{\tilde{y}}\left|\frac{\mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid X)}{\mathbb{P}^*(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid X)} - 1\right|, \max_{\tilde{y}}\left|\frac{\mathbb{P}^*(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid X)}{\mathbb{P}_\theta(A_{1:K} \in \mathcal{N}(\tilde{y}, r) \mid X)} - 1\right|\right\}\right] \\ + \frac{2}{\gamma} d_{\text{TV}}\left(\mathbb{P}_w(Y, A_{1:K}), \mathbb{P}^*(Y, A_{1:K})\right),\end{aligned}$$

where P^* represents the ground-truth distribution.

Proof. See proof for Theorem 6. □