# Learning Reach-Set Geometry for Tighter Probabilistic Neural Network Verification

**Samuel Sasaki** **Ben Wooding** **Taylor T. Johnson**

Institute for Convergent Software Integrated Systems
Vanderbilt University, USA
{samuel.sasaki, ben.wooding, taylor.johnson}@vanderbilt.edu

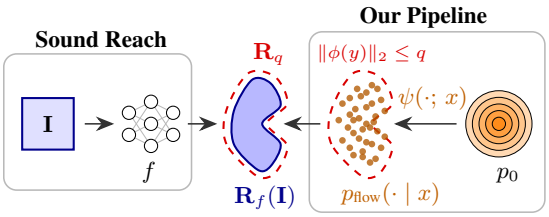
## Abstract

Neural network verification is key to certifying robustness in safety-critical systems. Sound verifiers commit to abstract domains that propagate soundly through the network's operations, paying for that commitment with looseness at scale and applicability largely confined to piecewise-linear architectures. Probabilistic verifiers relax soundness, but they too commit to a fixed family of shapes for the network's reachable outputs. When the network's true output geometry differs from the chosen shape, the verifier over-approximates the reach set and fails to certify networks that are in fact safe. We propose instead to *learn* the reach-set geometry from the network's behavior. A flow-matching model trained on input-output samples and calibrated by conformal prediction yields a probabilistic reach set whose shape adapts to the true shape of the reach set rather than to an a priori choice. The resulting set has no closed-form description, so we reframe the specification check as a rare-event estimation problem on the learned distribution, restricted to the calibration's coverage region, and solve it with bounded adaptive multilevel splitting. The pipeline produces a probabilistic safety certificate combining the conformal coverage with a rare-event upper bound. Our approach performs comparably to sound verifiers on a subset of VNN-COMP 2025 benchmarks where sound verification is tractable and matches or exceeds fixed-shape probabilistic baselines on most benchmarks in the same suite. On benchmarks whose perturbation model is a composition of geometric and photometric transformations rather than a small norm ball, the reach set turns non-convex, and we show that the learned geometry certifies far more than a calibrated convex baseline while sound verifiers run out of memory. On a synthetic family of growing-depth networks, it decides every instance with runtime growing only through inference cost where sound verifiers time out or return unknown. Code is available at `https://github.com/sammsaski/n2v-neurips2026`.

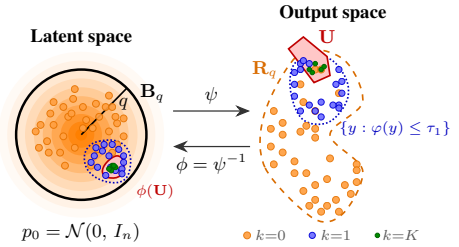
## 1 Introduction

Neural networks (NNs) are increasingly deployed in safety-critical roles, including perception and control in autonomous vehicles, learning-enabled controllers for cyber-physical systems, and decision-making in aerospace applications such as collision avoidance [Julian et al., 2016, Katz et al., 2017]. Because of the existence of adversarial perturbations [Goodfellow et al., 2015], we cannot reliably use NNs in these kinds of settings freely. Neural network verification addresses these concerns by checking whether the outputs of a given NN $f$ from a specified input region $\mathbf{I}$ avoid an unsafe region $\mathbf{U}$, i.e., checking $f(\mathbf{I}) \cap \mathbf{U} = \emptyset$. For example, local robustness checks that, given an input, we get the expected output regardless of perturbations to the original input up to some $\epsilon$.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).




(a) A comparison of two reachability approaches. *Sound reachability* (left) propagates the input set $\mathbf{I}$ through $f$ to the true reach set $\mathbf{R}_f(\mathbf{I})$ (blue). *Our pipeline* (right) trains a flow $\psi$ on $(x, f(x))$ pairs to map the standard Gaussian $p_0$ to an approximation $p_{\text{flow}}$ of $f$'s outputs, then conformally calibrates the latent norm at threshold $q$ to obtain the probabilistic reach set $\mathbf{R}_q$ (dashed red).




(b) Verifying a specification with bounded AMLS. A sequential MCMC chain in the latent space (left), restricted to the conformal ball $\mathbf{B}_q$, drives particles from a uniform distribution at $k{=}0$ through intermediate level sets ($k{=}1$) toward $\phi(\mathbf{U})$ at $k{=}K$. The flow $\psi$ pushes the partition into output space (right).

Figure 1: The two components of our proposed approach, learning the reach-set geometry with flow matching (a) and verifying the specification with bounded AMLS (b).

Sound verifiers provide formal guarantees about whether or not the specification is satisfied. Linear bound propagation with branch-and-bound [Xu et al., 2021, Wang et al., 2021], search-based methods such as NeuralSAT [Duong et al., 2025], and set-based reachability tools such as CORA [Althoff, 2015] and PyRAT [Lemesle et al., 2026], all top finishers at the Verification of Neural Networks Competition (VNN-COMP) [Kaulen et al., 2025], have driven substantial progress. However, their applicability is constrained by scaling and architectural challenges. The computational cost of case-splitting, bound propagation, and constraint solving scales poorly with parameter count, depth, and input-set size. The largest networks reliably handled by state-of-the-art neural network verification tools sit in the low millions of parameters, and even ResNet-50, a widely-deployed 25M-parameter image classifier, remains beyond the practical reach of these verifiers. Sound verification is also restricted in practice to piecewise-linear networks, where ReLU activations and affine layers admit tractable symbolic analysis. Attention, normalization layers, and smooth activations such as GELU and SiLU introduce nonlinearities for which sound verification is fundamentally far more difficult. As a consequence, the kinds of NNs driving current production systems, including vision-language-action models, large language models, and latent-diffusion image generators, lie beyond the reach of sound verification, even setting aside the scale gap.

These constraints have motivated a probabilistic relaxation of the verification problem. Probabilistic verification provides statistical guarantees that the network satisfies its specification with high probability under an assumed input distribution, trading the deterministic worst-case guarantees of sound verification for statistical ones in exchange for a pathway to more scalable approaches. Existing probabilistic verifiers nevertheless commit in advance to a fixed family of shapes for the output reachable set (reach set), including hand-designed nonconformity scores [Hashemi et al., 2025b], star-set propagation [Tran et al., 2025], parametric scenario regions [Lin and Bansal, 2024], and pointwise robustness certificates [Cohen et al., 2019]. Sound verifiers commit similarly through their choice of abstract domain. Whenever the chosen shape mismatches the network's true reach set, or the over-approximation grows loose at scale, the verifier overestimates and can fail to certify networks that are in fact safe. We instead ask whether the reach-set geometry can be *learned* from the network. Doing so reframes the verification challenge, since a learned reach set has no closed-form description, and standard specification-checking machinery either does not apply or falls back to a coarse over-approximation that discards the geometric tightness the learning step was meant to provide.

Our contributions are as follows:

**1) Geometry-aware probabilistic reach sets via flow matching.** We introduce the first conformal probabilistic reach set whose geometry is *learned* from the network's input-output behavior, using a flow-matching model's reverse-map latent norm as the nonconformity score. Existing conformal NN verification scores fix the reach-set geometry a priori while ours adapts to whatever shape $f$ induces.

**2) Bounded adaptive multilevel splitting for verification on the learned set.** We frame specification verification on the resulting non-convex reach set as a rare-event estimation problem and adapt adaptive multilevel splitting (AMLS) to operate inside the conformal coverage region, improving accuracy by eliminating spurious out-of-coverage counterexamples. Combined with the coverage guarantee, this yields a multiplicative joint safety certificate.

**3) Empirical validation.** On a subset of VNN-COMP benchmarks, our approach performs similarly to the 2025 winner and matches or exceeds the fixed-shape probabilistic baselines Clipping-Block [Hashemi et al., 2025a] and SAVER [Sivaramakrishnan et al., 2025] on most benchmarks. Under composed geometric and photometric perturbations, we certify 27 of 31 ResNet-50 instances at top-1 and all 42 at top-5, a strict superset of the 16 and 32 Clipping-Block certifies, and 80.3% of audited BiSeNet segmentation pixels against its 58.8%, while $\alpha$,$\beta$-CROWN runs out of memory on both networks. On a synthetic family of growing-depth networks, our approach decides every instance with runtime growing only via inference cost where sound verifiers time out or return unknown.

## 2 Preliminaries

**Reachability analysis for NNs.** We consider a deterministic feedforward neural network $f : \mathbb{R}^{n_0} \to \mathbb{R}^n$, with input dimension $n_0$ and output dimension $n$, and an input set $\mathbf{I} \subset \mathbb{R}^{n_0}$, typically an $\ell_\infty$-ball around a nominal input. The reach set under $f$ is

$$\mathbf{R}_f(\mathbf{I}) \;:=\; f(\mathbf{I}) \;=\; \{f(x) : x \in \mathbf{I}\} \;\subseteq\; \mathbb{R}^n, \tag{1}$$

the set of all outputs the NN can produce on $\mathbf{I}$. The standard neural network verification paradigm poses the problem of verifying the specification for a NN as a search for a counterexample, an input $x \in \mathbf{I}$ such that $f(x) \in \mathbf{U}$. Neural network verification via reachability analysis, then, certifies some specification for a NN by showing that no counterexample exists, equivalently $\mathbf{R}_f(\mathbf{I}) \cap \mathbf{U} = \emptyset$. For ReLU NNs, those which have ReLU activations and the general focus of the neural network verification community, computing $\mathbf{R}_f(\mathbf{I})$ exactly is generally intractable because the number of activation patterns is exponential in the number of ReLU neurons, and exact reach-set computation requires enumerating all activation regions. Sound reachability methods therefore construct an over-approximation $\widehat{\mathbf{R}}_f(\mathbf{I}) \supseteq \mathbf{R}_f(\mathbf{I})$, using abstractions like linear bounds [Tiwari, 2003], zonotopes [Girard, 2005], polyhedra [Tran et al., 2019b], or star sets [Tran et al., 2019a], and decide safety by checking $\widehat{\mathbf{R}}_f(\mathbf{I}) \cap \mathbf{U} = \emptyset$. Looseness in $\widehat{\mathbf{R}}_f(\mathbf{I})$ relative to $\mathbf{R}_f(\mathbf{I})$ produces *unknown* verdicts, where the verifier neither proves safety on the over-approximation nor exhibits a concrete counterexample on $f$ itself. Developing methods that construct $\widehat{\mathbf{R}}_f(\mathbf{I})$ as increasingly tighter over-approximations of $\mathbf{R}_f(\mathbf{I})$ at the scale of modern NNs is therefore critical for the safe deployment of neural networks in autonomous systems.

**Probabilistic verification** introduces an input distribution $\mathcal{W}$ on the input region $\mathbf{I}$, written $x \overset{\mathcal{W}}{\sim} \mathbf{I}$, and replaces the exact reach set $\mathbf{R}_f(\mathbf{I})$ with a probabilistic reach set that covers the true reach set $f(x)$ with a stated probability over $\mathcal{W}$. Conformal prediction [Vovk et al., 2005, Tibshirani et al., 2019] supplies the standard machinery. Given i.i.d. nonconformity scores $\{R_i\}_{i=1}^m$ drawn from a target distribution $\mathcal{D}$, a miscoverage level $\epsilon \in (0,1)$, and a rank $\ell \le m$, the order statistic $R_\ell$ satisfies the marginal coverage statement $\Pr[R_{m+1} \le R_\ell] \ge 1-\epsilon$, where the probability is taken over the joint randomness of the calibration sample and a fresh score.

We adopt the *double-step* refinement of Hashemi et al. [2025b], parameterized by the triple $\langle \epsilon, \ell, m \rangle$ with $\epsilon$ the miscoverage level, $m$ the calibration sample size, and $\ell$ the order-statistic rank used for the threshold. The double-step separates calibration confidence from coverage,

$$\Pr[\Pr[R^{\text{unseen}} \le R_\ell] > 1-\epsilon] \;>\; 1 - \mathsf{betacdf}_{1-\epsilon}(\ell,\, m+1-\ell). \tag{2}$$

This identity requires $\mathcal{D}$ to have a continuous CDF so that calibration-sample ties have zero probability. For the flow-matching score $\|\phi(\cdot)\|_2$ in Section 3, continuity follows from $\psi$ being a homeomorphism of $\mathbb{R}^n$ and $\mathcal{W}$ being continuous. The inner probability is the *coverage level* $\delta_1 := 1-\epsilon$, the chance that an unseen score lies below $R_\ell$. The outer probability is the *confidence level* $\delta_2 := 1 - \mathsf{betacdf}_{1-\epsilon}(\ell, m+1-\ell)$, the chance over the calibration sample that the coverage statement actually holds. For example, $\epsilon, \ell, m$ chosen such that $\delta_1 = 0.999$ and $\delta_2 = 0.997$ guarantee that the calibrated reach set contains a fresh output with probability at least $99.9\%$, and that this containment statement itself holds with confidence at least $99.7\%$ over the calibration draw.

For a NN $f$, an input distribution $\mathcal{W}$ on $\mathbf{I}$, and a nonconformity score $s : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ on outputs, drawing a calibration sample $\{x_i\}_{i=1}^m$ with $x_i \overset{\mathcal{W}}{\sim} \mathbf{I}$ and setting $R_i := s(f(x_i))$ yields the threshold $q := R_\ell$ and the *probabilistic reach set*

$$\mathbf{R}_f^\epsilon(\mathcal{W};\, \ell,\, m) \;:=\; \{y \in \mathbb{R}^n : s(y) \leq q\}, \tag{3}$$

which carries the $\langle \epsilon, \ell, m \rangle$ guarantee. We abbreviate this set as $\mathbf{R}_q$ from Section 3 onward whenever the dependence on $(\mathcal{W}, \ell, m)$ is fixed by context. Existing probabilistic neural network verification methods that follow this framework [Hashemi et al., 2025b,a] pick $s$ from a fixed family, hyperrectangle distance, clipping-block distance, and similar, so the geometry of $\mathbf{R}_q$ is determined a priori. Section 3 replaces $s$ with a learned score shaped by the network's input-output behavior on $\mathbf{I}$, so that $\mathbf{R}_q$ adapts to the NN's actual reachable geometry rather than being constrained to a shape chosen a priori.

**Flow matching.** Our reach-set construction relies on flow matching [Lipman et al., 2023], which learns a velocity field $u_\theta : \mathbb{R}^n \times [0,1] \to \mathbb{R}^n$ via the ODE

$$\frac{d}{dt}\,\psi_t(z) \;=\; u_\theta(\psi_t(z),\, t)\,, \qquad \psi_0(z) = z, \tag{4}$$

so that $\psi_1$ transports a base distribution $p_0 = \mathcal{N}(0, I_n)$ to a target $p_1$ on $\mathbb{R}^n$. For brevity we write $\psi := \psi_1$ for the time-1 transport map and $\phi := \psi^{-1}$ for its inverse. In our setting $p_1$ is the distribution of $f$'s outputs over $\mathbf{I}$, so the velocity network conditions on the network's input as a third argument, $u_\theta(y, t, x)$. Section 3 uses $\psi$ and $\phi$ to mimic reachability analysis with flow matching, where a flow trained on $(x, f(x))$ pairs gives a learned generative model of the network's reachable outputs and the reverse map's latent norm becomes the nonconformity score whose level sets define the probabilistic reach set $\mathbf{R}_q$.

We choose flow matching because the latent-norm score requires an exact deterministic inverse, and not every generative family supplies one. Conformal calibration admits any nonconformity score that is a fixed deterministic function of the output, so calibration scores stay exchangeable with a fresh draw, with a continuous distribution so the coverage statement holds without ties, and the exact inverse is what makes $\|\phi(y)\|_2$ such a function. GANs provide no inverse and VAE encoders only an approximate one that already breaks this exchangeability, while diffusion's probability-flow ODE is deterministic and closely related to flow matching, which we prefer for its simulation-free training objective [Lipman et al., 2023, Tong et al., 2024]. The latent norm is also structurally meaningful, since $\phi$ is a homeomorphism and the score's sub-level sets are images of Euclidean balls, making the induced reach sets connected, nested in the threshold, and ordered by base-measure density, structure an arbitrary admissible score does not provide.

## 3 Flow Matching-based Probabilistic Neural Network Verification

### 3.1 Flow matching for reach-set learning

Existing nonconformity scores in conformal NN verification impose a geometric form on $\mathbf{R}_q$ a priori [Hashemi et al., 2025b,a]. We replace the fixed score with one learned from $f$'s input-output behavior on $\mathbf{I}$.

**Training the flow.** We sample $x_1, \ldots, x_N \overset{\text{i.i.d.}}{\sim} \mathcal{W}$ on $\mathbf{I}$, evaluate $y_i = f(x_i)$ to build input-output pairs $\{(x_i, y_i)\}$ from the true reach set, and train a conditional OT-CFM model [Tong et al., 2024] on them following the setup of Section 2. The trained flow $\psi(\cdot;\, x)$ is invertible by construction and transports the standard Gaussian $p_0 = \mathcal{N}(0, I_n)$ to an approximation $p_{\text{flow}}(\cdot \mid x)$ of the input-conditional output distribution $p(\cdot \mid x)$ induced by $f$. The flow does not learn $f$ itself, only the distribution of $f$'s outputs under $\mathcal{W}$.

**Score construction.** The trained flow gives us a learned generative model of $f$'s outputs, and we want to convert it into a nonconformity score. We ask how typical a given output $y$ is under the learned distribution $p_{\text{flow}}(\cdot \mid x)$. The base distribution $p_0 = \mathcal{N}(0, I_n)$ assigns its mass near the origin and decays radially in the latent norm, so an output that is typical under $p_{\text{flow}}$ pulls back to a latent near the origin, while an output in the tail pulls back to a latent with large norm. The flow's reverse

map $\phi(\cdot;\, x)$ performs exactly this pull-back, integrating the ODE backward from $y$ to recover the latent $\phi(y;\, x)$ that the forward flow would have transported to $y$. The atypicality of $y$ is then captured by the latent norm,

$$s(y;\, x) \;:=\; \|\phi(y;\, x)\|_2. \tag{5}$$

Because the flow's coordinate system was shaped by $f$'s input-output behavior on $\mathbf{I}$, the level set $\{y : s(y;\, x) \le q\}$ adapts to the network's reachable geometry, fitting non-convexity in the true reach set rather than approximating it with a fixed shape. Figure 3 illustrates the refinement of the learned reach set to the true reach-set geometry through the training process on a synthetic example.

### 3.2 Conformal calibration of the reach set

With the score from (5) in hand, we instantiate the double-step calibration of Section 2. Drawing a fresh i.i.d. calibration sample $\{x_i\}_{i=1}^m$ with $x_i \overset{\mathcal{W}}{\sim} \mathbf{I}$, drawn independently of the flow-training sample so that the trained $\psi$ is fixed before calibration, we evaluate $y_i = f(x_i)$ and compute scores $R_i := s(y_i;\, x_i)$. The threshold $q := R_\ell$ is set as the $\ell$-th order statistic. Plugging (5) into (3), the general equation for a probabilistic reach set, gives the definition of a probabilistic reach set with the score as the latent norm under our trained flow:

$$\mathbf{R}_q \;:=\; \{y \in \mathbb{R}^n : \|\phi(y;\, x)\|_2 \le q\}. \tag{6}$$

The $\langle \epsilon, \ell, m \rangle$ guarantee carries over unchanged, $\Pr_{x\sim\mathcal{W}}[f(x) \in \mathbf{R}_q] \ge 1 - \epsilon$ with confidence at least $\delta_2 = 1 - \mathsf{betacdf}_{1-\epsilon}(\ell, m+1-\ell)$ over the calibration sample. The construction of $\mathbf{R}_q$ via a learned flow is what makes the geometry adaptive. Its flexibility, however, is also what makes $\mathbf{R}_q$ hard to verify a specification against, as we discuss next.

### 3.3 Specification verification via bounded adaptive multilevel splitting (AMLS)

Given the calibrated reach set $\mathbf{R}_q$ from (6), the verification question is whether $\mathbf{R}_q \cap \mathbf{U} = \emptyset$. For a closed-form $\mathbf{R}_q$, an axis-aligned box, an ellipsoid, or a polyhedron, this reduces to a closed-form geometric check using standard techniques such as intersecting halfspaces, projecting onto axes, or solving a small linear program. Our $\mathbf{R}_q$ is not closed-form. The forward flow $\psi(\cdot;\, x)$ is a learned ODE-defined homeomorphism with no analytical inverse beyond numerical ODE integration, and $\mathbf{R}_q$ is non-convex in general, so the only operation we can perform on $\mathbf{R}_q$ is a membership test. A naive workaround that replaces $\mathbf{R}_q$ with a convex over-approximation would erase the geometric tightness that the learned reach set was constructed to provide. The verification question must therefore be answered by sampling from $\mathbf{R}_q$ and checking whether sampled points fall in $\mathbf{U}$, rather than by inspecting the geometry of $\mathbf{R}_q$ directly.

**Latent-space equivalence.** Because $\phi$ and $\psi$ are inverse homeomorphisms, $\mathbf{R}_q$ corresponds under the change of coordinates $z = \phi(y;\, x)$ to the latent Euclidean ball $\mathbf{B}_q := \{z \in \mathbb{R}^n : \|z\|_2 \le q\}$. Verifying $\mathbf{R}_q \cap \mathbf{U} = \emptyset$ is equivalent to verifying $\mathbf{B}_q \cap \psi^{-1}(\mathbf{U};\, x) = \emptyset$, with sampling under $\pi_q := p_0(\cdot \mid \|z\|_2 \le q)$, the standard Gaussian truncated to $\mathbf{B}_q$.

**Mass formulation.** The verification question reduces to estimating one probability. How much of the flow's mass on the conformal ball $\mathbf{B}_q$ lands in the unsafe region after pushing through the flow?

$$\pi \;:=\; \Pr_{z\sim\pi_q}[\psi(z;\, x) \in \mathbf{U}] \;=\; \Pr_{y\sim p_{\text{flow}}}[y \in \mathbf{U} \mid y \in \mathbf{R}_q]. \tag{7}$$

This is the in-coverage unsafe-region mass under the flow's pushforward. A high-confidence upper bound $\hat{\pi}_{\text{upper}} \le \eta$ at target tail level $\eta$ certifies safety against $\mathbf{U}$ on $\mathbf{R}_q$.

**Adaptive multilevel splitting.** If $\pi$ is small (say $10^{-6}$), drawing $N$ i.i.d. samples and counting how many fall in $\mathbf{U}$ gives a useful upper bound only for $N \gg 1/\pi$, which is infeasible at deployment scale. Adaptive multilevel splitting (AMLS) [Cérou and Guyader, 2007, Cérou et al., 2019] sidesteps this by replacing the single rare event with a sequence of easier ones. Pick an unsafe-region margin function $\varphi$ with $\varphi(y) \le 0 \Leftrightarrow y \in \mathbf{U}$, and a sequence of decreasing thresholds $\tau_0 > \tau_1 > \cdots > \tau_K = 0$. Then $\pi$ factorizes as

$$\pi \;=\; \prod_{k=1}^{K} \Pr\big[\varphi(\psi(z;\, x)) \le \tau_k \;\big|\; \varphi(\psi(z;\, x)) \le \tau_{k-1}\big]\,, \tag{8}$$

a product of conditional probabilities, each at a moderate target rate $\rho$ that we control. AMLS estimates each factor by simulating a Markov chain conditioned on the previous threshold, and chooses each $\tau_k$ adaptively from the chain's quantiles. The Cérou-Guyader CLT [Cérou and Guyader, 2007] supplies an asymptotic upper bound $\hat{\pi}_{\text{upper}}$ at confidence $1 - \delta_3$. AMLS for neural network reliability has prior work [Webb et al., 2018, TIT et al., 2023], but it samples from $p_0$ on all of $\mathbb{R}^n$, which is not the conformal coverage region we need.

**Bounded AMLS.** Without modification, the chain can drift outside $\mathbf{B}_q$. An unsafe latent found there does not say anything about $\mathbf{R}_q$, since the conformal coverage guarantee is only over the in-ball region. Treating an out-of-ball latent as evidence of unsafety would over-reject safe specifications. We restrict the chain to $\|z\|_2 \leq q$ at every state. Under this restriction, $\hat{\pi}_{\text{upper}}$ tracks the in-ball mass $\pi$ from Equation (7) and the rare-event search domain matches the conformal coverage region exactly. The restriction adds a single in-ball check at each chain step, and we adopt the CLT transfer of Cérou and Guyader [2016] to the in-ball-restricted chain as Assumption 2 (Appendix B).

The latent-norm score and this bounded search are co-designed. The score's conformal region *is* the ball $\mathbf{B}_q$ the chain searches, so the in-ball check costs one norm comparison, whereas the flow's log-density, the natural alternative score, yields a region whose latent preimage is not a ball, turning every membership test into a divergence-integrated ODE solve and breaking the truncated-Gaussian structure behind Assumption 2. A comparison against the log-density score on geometries with known ground-truth reach sets shows no loss in tightness (Appendix B).

### 3.4 Joint guarantee

Combining the conformal coverage of $\mathbf{R}_q$ from Section 3.2 with the bounded-AMLS upper bound on the in-ball unsafe mass from Section 3.3 yields the paper's main probabilistic guarantee.

**Theorem 1** (Joint coverage + tail bound)**.** Let $\mathbf{R}_q$ be the probabilistic reach set obtained by double-step calibration of the latent norm $s(y;\, x) = \|\phi(y;\, x)\|_2$ at miscoverage level $\epsilon$, calibration size $m$, and rank $\ell$, with calibration confidence $\delta_2 = 1 - \texttt{betacdf}_{1-\epsilon}(\ell, m + 1 - \ell)$. Let $\hat{\pi}_{\text{upper}}$ be the $1 - \delta_3$ asymptotic upper bound on the in-ball unsafe mass $\pi$ obtained by bounded AMLS under Assumption 2, and assume the flow's conditional pushforward agrees with the true input-conditional output distribution on $\mathbf{R}_q$ (the flow-fidelity assumption discussed in Section 6). If the bounded-AMLS chain reports no in-ball detection and $\hat{\pi}_{\text{upper}} \leq \eta$, then with confidence at least $\delta_2(1 - \delta_3)$ over the joint randomness of the calibration and AMLS samples,

$$\Pr_{x \sim \mathcal{W}}[f(x) \in \mathbf{U}] \;\leq\; 1 - (1 - \epsilon)(1 - \eta) \;\leq\; \epsilon + \eta. \tag{9}$$

*Proof.* See Appendix A. □

At our operating point ($\epsilon = 10^{-3}$, $\eta = 10^{-4}$, $m = 8000$, $\ell = 7999$, $\delta_3 = 0.001$), the joint bound is $\approx 1.1 \times 10^{-3}$ at confidence $\delta_2(1 - \delta_3) \approx 0.996$ over the calibration and AMLS draws. The guarantee is conditional on the flow-fidelity assumption, which no trained flow satisfies exactly, so we read Theorem 1 as an idealization. Theorem 4 in Appendix F replaces the assumption with a measurable gap, a conditional total-variation distance of at most $\tau$ between the flow and the true output distribution inside $\mathbf{R}_q$, and certifies the weaker bound $\epsilon + \eta + \tau$. Theorem 1 is the $\tau = 0$ case. Flow error therefore enters the certificate only through the additive term $\tau$. The coverage statement is unaffected, since it is finite-sample and holds for any fixed score.

## 4 Evaluation

### 4.1 Experimental setup

We use $\epsilon = 10^{-3}$, $\eta = 10^{-4}$, $m = 8000$, and $\ell = 7999$ across all runs, with per-benchmark training budgets and AMLS knob settings reported in Appendix D. Probabilistic baselines lacking a built-in falsifier are paired with a PGD/APGD ensemble on $f$ for parity with VNN-COMP verdicts. We report verdict counts (SAT, UNSAT, unknown, timeout), UNSAT recall, false-UNSAT rate, and wall-clock. All experiments were run on a single A30 GPU with 24 GB of VRAM.

Table 1: UNSAT accuracy (%) on VNN-COMP'25 benchmarks, with the full % Solved view in Appendix G. A dagger ($\dagger N$) marks $N$ sound violations, UNSAT verdicts whose ground truth is SAT.

| **Method** | ACAS Xu | collins rul-cnn | dist shift | linear-izenn | tll-verify | meta-room | vit | tiny imagenet | cifar 100 |
|---|---|---|---|---|---|---|---|---|---|
| **Ours** | **96.4%**$^{\dagger 1}$ | **97.5%** | **95.5%** | 90.0% | **93.8%** | **99.0%**$^{\dagger 2}$ | **99.5%** | **99.4%**$^{\dagger 4}$ | **98.8%**$^{\dagger 3}$ |
| *Sound verifiers (from VNN-COMP'25)* | | | | | | | | | |
| $\alpha\beta$-CROWN | **100.0%** | **100.0%** | **100.0%** | **100.0%** | **100.0%** | **99.0%** | **62.7%** | **84.7%** | **58.7%** |
| NeuralSAT | 100.0% | 100.0% | 100.0% | 100.0% | 100.0% | 99.0% | 33.8% | 71.8% | 51.2% |
| PyRAT | 100.0% | 100.0% | 98.5% | 100.0% | 100.0% | 100.0%$^{\dagger 2}$ | 41.8% | 42.3% | 36.0% |
| CORA | 98.6% | 100.0% | 100.0% | 100.0% | 100.0% | 96.9% | – | 0.6% | 0.6% |
| *Probabilistic verifiers* | | | | | | | | | |
| Clipping-Block | 82.9% | 95.0% | 89.4% | **98.3%** | **93.8%** | 92.7% | **99.5%** | **99.4%**$^{\dagger 4}$ | **99.4%**$^{\dagger 3}$ |
| SAVER | 98.6%$^{\dagger 41}$ | 97.5% | 98.5%$^{\dagger 4}$ | 98.3%$^{\dagger 1}$ | 93.8%$^{\dagger 2}$ | 0.0% | 0.0% | 0.0% | 0.0% |

**Verification benchmarks.** Our benchmarks fall into two groups, with the first being nine benchmarks from VNN-COMP 2025. Six are ACAS Xu [Julian et al., 2016, Katz et al., 2017], the Collins RUL 1D CNN regression task, and the image classifiers dist-shift, linearizenn, tllverify, and metaroom. The remaining three (a vision transformer, a TinyImageNet classifier [Le and Yang, 2015], and a CIFAR-100 classifier [Krizhevsky, 2009], all with multi-class robustness specifications) are high-output-dimension classifiers where published sound verifiers fail to reach a verification decision within the allotted time on a significant number of instances. Results for sound verifiers are taken from VNN-COMP 2025. The probabilistic methods are run at each instance's published timeout.

The second group goes beyond the perturbation models of the first. Those input sets are mostly small $\ell_\infty$ balls, under which a calibrated convex shape already fits the reachable outputs tightly, so they cannot separate learned geometry from a well-chosen fixed shape. We therefore add two tasks whose input sets come from the semantic-perturbation literature [Balunovic et al., 2019, Wang et al., 2023]. The first task verifies top-1 and top-5 robustness of ResNet-50 [He et al., 2016] on 50 ImageNet images. In place of a pixel-space ball, the input set $\mathbf{I} \subset \mathbb{R}^{11}$ is a box of transformation parameters around the identity, spanning rotation $\pm 5^\circ$, scale $\pm 2\%$, per-axis translation $\pm 2\%$, brightness $\pm 0.02$, contrast $\pm 3\%$, gamma $\pm 3\%$, saturation $\pm 5\%$, and per-channel RGB gains $\pm 2\%$. The distribution $\mathcal{W}$ is uniform over $\mathbf{I}$, and the verified map sends $\theta \in \mathbf{I}$ to the ResNet-50 logits of the image transformed by $\theta$, a setting we refer to as the composed transformation model. The second task verifies per-pixel argmax robustness of the BiSeNet CamVid segmentation model [Yu et al., 2018] released with Clipping-Block [Hashemi et al., 2025a], with $\mathbf{I} \subset \mathbb{R}^7$ the photometric sub-box at twice the half-widths above, on 2,136 interior pixels across 25 images. Both methods receive identical input samples, calibration budget ($m = 8000$), falsifier, and network-evaluation budget at comparable confidence ($0.996$ ours, $0.997$ Clipping-Block), with Clipping-Block run as proposed, a PCA-reduced sampled hull with per-coordinate conformal inflation. For these two tasks our chain additionally abstains from certifying below a margin threshold fixed before evaluation, and every certificate it issues is re-checked against fresh input draws, with the full protocol in Appendix E.

**Volume comparison.** We measure how tightly each method wraps the true reach set on seven synthetic benchmarks where the ground-truth reach-set volume is known. Identity-activation 1-Lipschitz networks at output dimensions $n \in \{2, 3, 5, 10, 20\}$ admit a closed-form volume, and 2D and 3D banana-shaped benchmarks have non-convex reach sets whose volume we estimate by high-sample Monte Carlo. We compare three methods (our flow-based reach set, Clipping-Block, and reachability analysis with star-set over-approximation) at three calibration sample budgets, $m \in \{1\text{k}, 8\text{k}, 16\text{k}\}$ with 5 seeds each. The metric is the volume ratio $\mathrm{vol}(\mathbf{R}_{\text{predicted}})/\mathrm{vol}(\mathbf{R}_{\text{true}})$, where smaller is tighter and $1.00$ is exact.

**Controlled scaling.** We isolate how each method's wall-clock cost grows with network size by varying depth while fixing every other aspect of the network and the specification. We construct a family of ReLU MLPs at depths $d \in \{2, 4, 8, 16, 24, 32, 40\}$ with width $512$ and spectrally-normalized linear layers, hence 1-Lipschitz by composition, paired with a halfspace specification beyond the empirical maximum of $f(\mathbf{I})$ so that every instance is UNSAT by Lipschitz construction.

Table 2: Certification beyond convex-friendly perturbation models. $n$ counts instances with no known counterexample and percentages are over that pool. The last column (ours / Clipping-Block) counts certificates issued on falsified instances, reported as SAT under falsifier pairing. $\alpha$,$\beta$-CROWN runs out of memory on both models.

| Model | Specification | $n$ | Ours | Clipping-Block | False UNSATs |
|---|---|---|---|---|---|
| ResNet-50 | top-1 | 31 | **27** (87.1%) | 16 (51.6%) | 1 / 0 |
| ResNet-50 | top-5 | 42 | **42** (100%) | 32 (76.2%) | 1 / 0 |
| BiSeNet (segmentation) | per-pixel top-1 | 1,296 | **1,041** (80.3%) | 762 (58.8%) | 5 / 0 |

Each method's verdict therefore reduces to whether it finishes within the per-instance 300s budget. For this experiment, we compare our method against Clipping-Block, $\alpha$,$\beta$-CROWN, and NeuralSAT.

**Ablation of the specification-verification method.** We ablate verification methods on full ACAS Xu (186 instances) and tllverify (32 instances). To isolate verifier quality from randomness in the flow training process, we calibrate the flow once per instance and run all five approaches against the same $(\text{flow}, q)$ tuple. The five techniques are scenario sampling, AMLS-unbounded, AMLS-bounded (ours), importance sampling, and raw uniform Monte Carlo. We measure UNSAT recall on ground-truth-UNSAT instances and sound-violation count on ground-truth-SAT instances.

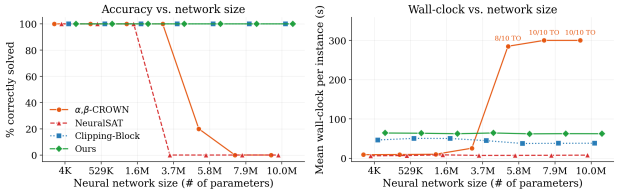
### 4.2 Discussion

**Verdict accuracy across VNN-COMP.** Because the pipeline certifies only UNSAT, Table 1 reports UNSAT accuracy across the VNN-COMP benchmarks (full verdict matrix in Appendix G). Our pipeline produces verdicts comparable to the VNN-COMP 2025 winner $\alpha$,$\beta$-CROWN where sound verification is tractable, and recovers more UNSAT verdicts than the probabilistic baselines Clipping-Block and SAVER on the same suite.

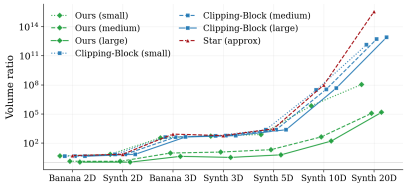
**Beyond convex-friendly perturbation models.** Table 2 reports results under the two composed perturbation models. On ResNet-50 we certify every instance Clipping-Block certifies plus 11 more at top-1 and 10 more at top-5, at 125 s per instance against 264 s, and on BiSeNet we certify 80.3% of audited pixels against 58.8%. The VNN-COMP benchmark suite cannot surface this gap because of how its input sets are built. Within a small $\ell_\infty$ ball the network responds approximately linearly, so the reachable outputs form a single ellipsoid-like cloud that is already nearly convex, a calibrated convex shape wraps that cloud almost exactly, and a learned geometry has nothing left to win. Consistent with this account, the two methods agree on 200 of 200 instances on an $\ell_\infty$-robust ResNet-50 (Appendix E). The transformation boxes are instead low-dimensional and map nonlinearly to pixels, so the reachable outputs curve, and a convex surrogate must inflate until it covers curvature it cannot represent, giving up every certificate whose specification margin that inflation consumes. The learned reach set follows the curvature instead of paying for it, the same behavior the volume study of Figure 2b measures against ground truth. Tightness that overclaims would be no gain, so every certificate the chain issued on ResNet-50 was re-checked against its certified bound on unsafe mass with fresh draws of the transformation parameters, and exactly one failed, which we report as an empirical certificate failure with the full audit in Appendix E.

**Sound-verifier failure modes.** Figure 2a surfaces sound-verifier failure modes that are intrinsic to the underlying technique rather than to implementation artifacts. $\alpha$,$\beta$-CROWN's branch-and-bound times out at the larger depths in our family because its branching cost grows rapidly once the network is deep enough that the linear bounds it relies on stop being tight. NeuralSAT's abstractor returns UNKNOWN at roughly constant wall-clock once its bounds are too loose to enter branch-and-bound, reporting that it cannot decide rather than timing out. Meanwhile, our approach and Clipping-Block decide every instance at a cost growing with inference time alone, since a sampling-based verifier needs only forward passes where a sound one must cover every behavior the network can realize.

**Geometry-aware tightness.** Figure 2b reports the volume ratio $\text{vol}(\mathbf{R}_{\text{predicted}})/\text{vol}(\mathbf{R}_{\text{true}})$ on synthetic benchmarks where the ground-truth reach-set volume is known in closed form or through high-sample Monte Carlo estimates. Our flow-based reach set tightens monotonically with calibration sample budget, recovering the true volume to within a small factor in low dimensions and a manage-




(a) Wall-clock time vs network depth on the controlled-scaling family of 1-Lipschitz networks, paired with a halfspace specification disjoint from $f(\mathbf{I})$ so every instance is UNSAT. $\alpha$,$\beta$-CROWN's branch-and-bound grows exponentially in depth and times out at the largest depths. NeuralSAT abstains (UNKNOWN at near-constant cost) once its abstractor's bounds become too loose to enter branch-and-bound. Clipping-Block and our pipeline hold full accuracy across the range with runtime growing far more slowly in depth.




(b) Volume ratio $\mathrm{vol}(\mathbf{R}_{\text{predicted}})/\mathrm{vol}(\mathbf{R}_{\text{true}})$ on synthetic benchmarks with computable ground-truth reach sets. Smaller is tighter and 1.00 is exact. The flow-based reach set tightens with calibration budget $m$, while the fixed-shape baselines plateau at the mismatch their families induce.

Figure 2: Sound-verifier scaling behavior (a) and reach-set tightness against known ground truth (b).

able factor as output dimension grows. Clipping-Block and the star-set over-approximation plateau at the volume their fixed shape families induce, with the gap growing rapidly with output dimension.

**Specification-verification results.** Table 3 ablates the specification verification methods on full ACAS Xu and tllverify, with calibration shared across primitives so flow randomness is held constant. Bounded AMLS is the only approach simultaneously tied for best accuracy and lowest sound-violation count on both benchmarks. It ties scenario sampling and tilted importance sampling at $97.1\%$ accuracy on ACAS Xu while incurring a quarter of their sound-violation count ($1/47$ vs. $4/47$), and it ties them at $100\%/0$ on tllverify while strictly dominating unbounded AMLS ($46.7\%/0$). Scenario sampling and tilted importance sampling have elevated false-UNSAT rates on ACAS Xu, raw uniform Monte Carlo lacks the budget to reach low-probability events, and unbounded AMLS finds spurious counterexamples outside the conformal coverage region on tllverify. Bounded AMLS resolves both failure modes by restricting the chain to $\mathbf{B}_q$, recovering the in-coverage unsafe-mass estimate that Theorem 1 requires.

Table 3: Ablation of sampling-based specification-verification methods. *Accuracy* ↑ is the fraction of effective-UNSAT instances correctly certified UNSAT (breakdown shown) and *Sound viol.* ↓ counts UNSAT verdicts on ground-truth SAT instances.

| | **ACAS Xu** | | **tllverify** | |
|---|---|---|---|---|
| **Method** | Accuracy ↑ | Sound viol. ↓ | Accuracy ↑ | Sound viol. ↓ |
| Scenario | **97.1% (136/140)** | 4/47 | **100.0% (15/15)** | **0/17** |
| AMLS (unbounded) | 96.4% (135/140) | **1/47** | 46.7% (7/15) | **0/17** |
| **AMLS-bounded** | **97.1% (136/140)** | **1/47** | **100.0% (15/15)** | **0/17** |
| IS (tilted) | **97.1% (136/140)** | 4/47 | **100.0% (15/15)** | **0/17** |
| Raw MC (uniform) | 0.0% (0/140) | 0/47 | 0.0% (0/15) | 0/17 |

**Icicle geometry of false UNSATs.** A small number of false UNSATs in our VNN-COMP results trace to the "icicle" failure mode of Campi and Garatti [2018], where the falsifying region $\{x \in \mathbf{I} : f(x) \in \mathbf{U}\}$ is non-empty but a thin, low-measure sliver under $\mathcal{W}$. Where sound verification must abstain on these slivers, our joint bound $1-(1-\epsilon)(1-\eta) \approx 1.1 \times 10^{-3}$ upper-bounds the sliver's measure under $\mathcal{W}$ directly, with per-benchmark empirical rates against this bound reported in Appendix H.

## 5 Related Work

**Sound verifiers** produce yes/no answers with mathematical certainty. The dominant lineage is linear bound propagation combined with branch-and-bound, with $\alpha$,$\beta$-CROWN [Xu et al., 2021,

Wang et al., 2021] leading recent VNN-COMP editions [Kaulen et al., 2025]. SMT- and MIP-based verifiers [Duong et al., 2025, Katz et al., 2019] encode the network and specification into a satisfiability or mixed-integer-programming instance, and polyhedral and star-set abstract-interpretation approaches [Lemesle et al., 2026, Althoff, 2015, Tran et al., 2020, Tumlin et al., 2026] propagate abstract domains through the network to decide safety on the resulting over-approximation. Closed-loop reachability extends this through the controlled dynamics for neural-network-controlled cyber-physical systems [Ivanov et al., 2019, Huang et al., 2022, Ivanov et al., 2021]. An adjacent line addresses geometric rather than norm-ball perturbations. DeepG [Balunovic et al., 2019] computes sound relaxations for compositions of transformations, and GeoRobust [Wang et al., 2023] and the Hölder-optimization method of Zhang et al. [2025] search for a worst-case transform by global optimization at scales sound relaxations do not reach. These methods answer the existential question of whether a violating transform exists rather than bounding unsafe probability. Sound verification's structural limitations are what probabilistic verification, including our work, attempts to relax.

**Probabilistic verifiers** trade certainty for scale but each pre-commits to a fixed reach-set representation. The closest predecessor of our work is Hashemi et al. [2025b], who calibrate a probabilistic reach set with a hand-chosen score (axis-aligned hyperrectangle or clipping-block) and a double-step refinement of the conformal quantile, a skeleton we keep while replacing the score with the latent norm of a learned flow (Section 3). Lin and Bansal [2024] establish an equivalence between split conformal prediction and robust scenario optimization for neural reachable tubes with a fixed pointwise score. ProbStar [Tran et al., 2025] propagates a star set with associated probability mass through the network, inheriting the scaling and architectural limits of star-set sound verification. SAVER [Sivaramakrishnan et al., 2025] certifies coverage of an isotropically-shifted user-supplied SDF level set via a Campi-Garatti-style scenario sample-complexity bound [Campi and Garatti, 2018, Devonport and Arcak, 2020], with the shape itself fixed a priori. Randomized smoothing [Cohen et al., 2019, Salman et al., 2020, Yang et al., 2020] certifies a per-input robust radius via Gaussian noise injection rather than a reach set over an input region. Verifiably Robust Conformal Prediction [Jeary et al., 2024] places a verifier inside the conformal procedure, complementary to our placement of a generative model there. Subset simulation has been applied to NN verification in input space [Webb et al., 2018, TIT et al., 2023], while bounded AMLS is the first instance of the method applied in the latent space of a learned flow with the truncation bound set by the conformal threshold itself. Classical rare-event estimators, importance sampling [Owen and Zhou, 2000, Webb et al., 2018] and latent-space Langevin or HMC [Roberts and Tweedie, 1996, Neal, 2011], do not transfer to the coverage-restricted regime.

**Flow matching** as a generative modeling family is established by Lipman et al. [2023], with the OT-CFM variant of Tong et al. [2024] that we adopt for training and the closely related rectified-flow construction [Liu et al., 2022]. Output-distribution estimation with VAEs, GANs, and normalizing flows has been explored in safety-adjacent contexts including uncertainty quantification [Berry and Meger, 2023, Shen et al., 2022] and OOD detection [Zisselman and Tamar, 2020, Kirichenko et al., 2020]. To our knowledge, this paper is the first use of flow matching to construct a probabilistic neural network reach set with a coverage guarantee.

## 6 Conclusion

We learn the reach-set geometry of a neural network via flow matching, conformally calibrate the latent norm of the reverse map, and verify specifications on the resulting non-convex set with bounded AMLS. On nine VNN-COMP'25 benchmarks our pipeline matches sound verifiers where they are tractable and matches or exceeds fixed-shape probabilistic baselines on most benchmarks, with empirical false-UNSAT rates under $2.5\%$. On ResNet-50 under composed transformations and on BiSeNet segmentation, where sound verifiers run out of memory, the learned geometry converts its tightness into certificates that fixed-shape methods cannot issue.

**Limitations and future work.** The guarantee is probabilistic per certificate, holding with confidence $\delta_2(1-\delta_3)$ rather than with certainty, so a fraction of issued certificates will be false and a property reported verified can in fact be violated. The pipeline also never returns *SAT*, the coverage statement holds only under the assumed input distribution $\mathcal{W}$, and Theorem 1 assumes exact flow fidelity. Future directions include a negative-aware training objective and a non-asymptotic finite-sample bound for the in-ball AMLS estimator that would discharge Assumption 2.

## Acknowledgments and Disclosure of Funding

The material presented in this paper is based upon work supported by the National Science Foundation (NSF) through grant number 2220401 and the Defense Advanced Research Projects Agency (DARPA) under contract number FA8750-23-C-0518. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of DARPA or NSF.

## References

M. Althoff. An introduction to CORA 2015. In *Proc. of the 1st and 2nd Workshop on Applied Verification for Continuous and Hybrid Systems*, pages 120–151. EasyChair, 2015. doi: 10.29007/zbkv. URL https://easychair.org/publications/paper/xMm.

A. N. Angelopoulos, S. Bates, E. J. Candès, M. I. Jordan, and L. Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641 – 1662, 2025. doi: 10.1214/24-AOAS1998. URL https://doi.org/10.1214/24-AOAS1998.

M. Balunovic, M. Baader, G. Singh, T. Gehr, and M. Vechev. Certifying geometric robustness of neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/f7fa6aca028e7ff4ef62d75ed025fe76-Paper.pdf.

L. Berry and D. Meger. Normalizing flow ensembles for rich aleatoric and epistemic uncertainty modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6806–6814, 2023.

M. C. Campi and S. Garatti. *Introduction to the scenario approach*. SIAM, 2018.

F. Cérou and A. Guyader. Adaptive multilevel splitting for rare event analysis. *Stochastic Analysis and Applications*, 25(2):417–443, 2007.

F. Cérou and A. Guyader. Fluctuation analysis of adaptive multilevel splitting. *The Annals of Applied Probability*, 26(6):3319 – 3380, 2016. doi: 10.1214/16-AAP1177. URL https://doi.org/10.1214/16-AAP1177.

F. Cérou, A. Guyader, and M. Rousset. Adaptive multilevel splitting: Historical perspective and recent results. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(4), 2019.

J. Cohen, E. Rosenfeld, and Z. Kolter. Certified adversarial robustness via randomized smoothing. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1310–1320. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/cohen19c.html.

A. Devonport and M. Arcak. Estimating reachable sets with scenario optimization. In *Learning for dynamics and control*, pages 75–84. PMLR, 2020.

H. Duong, T. Nguyen, and M. B. Dwyer. NeuralSAT: A High-Performance Verification Tool for Deep Neural Networks. In R. Piskac and Z. Rakamarić, editors, *Computer Aided Verification*, pages 409–423, Cham, 2025. Springer Nature Switzerland.

A. Girard. Reachability of uncertain linear systems using zonotopes. In *International workshop on hybrid systems: Computation and control*, pages 291–305. Springer, 2005.

I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples, 2015. URL https://arxiv.org/abs/1412.6572.

N. Hashemi, S. Sasaki, D. M. Lopez, L. Lindemann, I. Oguz, M. Ma, and T. T. Johnson. Probabilistic robustness analysis in high dimensional space: Application to semantic segmentation network, 2025a. URL https://arxiv.org/abs/2509.11838.

N. Hashemi, S. Sasaki, I. Oguz, M. Ma, and T. T. Johnson. Scaling data-driven probabilistic robustness analysis for semantic segmentation neural networks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=liefJOFVfH.

K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition . In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Los Alamitos, CA, USA, June 2016. IEEE Computer Society. doi: 10.1109/CVPR.2016.90. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.90.

C. Huang, J. Fan, X. Chen, W. Li, and Q. Zhu. Polar: A polynomial arithmetic framework for verifying neural-network controlled systems. In *International Symposium on Automated Technology for Verification and Analysis*, pages 414–430. Springer, 2022.

R. Ivanov, J. Weimer, R. Alur, G. J. Pappas, and I. Lee. Verisig: verifying safety properties of hybrid systems with neural network controllers. In *Proceedings of the 22nd ACM International Conference on Hybrid Systems: Computation and Control*, pages 169–178, 2019.

R. Ivanov, T. Carpenter, J. Weimer, R. Alur, G. Pappas, and I. Lee. Verisig 2.0: Verification of neural network controllers using taylor model preconditioning. In *International Conference on Computer Aided Verification*, pages 249–262. Springer, 2021.

L. Jeary, T. Kuipers, M. Hosseini, and N. Paoletti. Verifiably robust conformal prediction. *Advances in Neural Information Processing Systems*, 37:4295–4314, 2024.

K. D. Julian, J. Lopez, J. S. Brush, M. P. Owen, and M. J. Kochenderfer. Policy compression for aircraft collision avoidance systems. In *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*, pages 1–10, 2016. doi: 10.1109/DASC.2016.7778091.

G. Katz, C. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer. Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks. In R. Majumdar and V. Kunčak, editors, *Computer Aided Verification*, pages 97–117, Cham, 2017. Springer International Publishing. ISBN 978-3-319-63387-9.

G. Katz, D. A. Huang, D. Ibeling, K. Julian, C. Lazarus, R. Lim, P. Shah, S. Thakoor, H. Wu, A. Zeljić, et al. The marabou framework for verification and analysis of deep neural networks. In *International conference on computer aided verification*, pages 443–452. Springer, 2019.

K. Kaulen, T. Ladner, S. Bak, C. Brix, H. Duong, T. Flinkow, T. T. Johnson, L. Koller, E. Manino, T. H. Nguyen, and H. Wu. The 6th international verification of neural networks competition (vnn-comp 2025): Summary and results, 2025. URL https://arxiv.org/abs/2512.19007.

P. Kirichenko, P. Izmailov, and A. G. Wilson. Why normalizing flows fail to detect out-of-distribution data. *Advances in neural information processing systems*, 33:20578–20589, 2020.

A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Y. Le and X. Yang. Tiny ImageNet visual recognition challenge, 2015.

A. Lemesle, J. Lehmann, T. L. Gall, and Z. Chihani. Verifying Neural Networks with PyRAT. In H. Oh and Y. Sui, editors, *Static Analysis*, pages 11–33, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-07106-4.

A. Lin and S. Bansal. Verification of neural reachable tubes via scenario optimization and conformal prediction. In A. Abate, M. Cannon, K. Margellos, and A. Papachristodoulou, editors, *Proceedings of the 6th Annual Learning for Dynamics & Control Conference*, volume 242 of *Proceedings of Machine Learning Research*, pages 719–731. PMLR, 15–17 Jul 2024. URL https://proceedings.mlr.press/v242/lin24a.html.

Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

X. Liu, C. Gong, and Q. Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

R. M. Neal. MCMC using Hamiltonian dynamics. *Handbook of markov chain monte carlo*, pages 47–95, 2011.

A. Owen and Y. Zhou. Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449):135–143, 2000.

G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. 1996.

H. Salman, M. Sun, G. Yang, A. Kapoor, and J. Z. Kolter. Denoised smoothing: A provable defense for pretrained classifiers. *Advances in Neural Information Processing Systems*, 33:21945–21957, 2020.

J. Shen, A. Agudo, F. Moreno-Noguer, and A. Ruiz. Conditional-flow nerf: Accurate 3d modelling with reliable uncertainty quantification. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022.

V. Sivaramakrishnan, K. C. Kalagarla, R. Devonport, J. Pilipovsky, P. Tsiotras, and M. Oishi. SAVER: A Toolbox for SAmpling-Based, Probabilistic VERification of Neural Networks. In *Proceedings of the 28th ACM International Conference on Hybrid Systems: Computation and Control*, HSCC '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400715044. doi: 10.1145/3716863.3718045. URL https://doi.org/10.1145/3716863.3718045.

R. J. Tibshirani, R. F. Barber, E. J. Candès, and A. Ramdas. *Conformal prediction under covariate shift*. Curran Associates Inc., Red Hook, NY, USA, 2019.

K. TIT, T. Furon, and M. Rousset. Gradient-informed neural network statistical robustness estimation. In F. Ruiz, J. Dy, and J.-W. van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 323–334. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/tit23a.html.

A. Tiwari. Approximate reachability for linear systems. In *International Workshop on Hybrid Systems: Computation and Control*, pages 514–525. Springer, 2003.

A. Tong, K. FATRAS, N. Malkin, G. Huguet, Y. Zhang, J. Rector-Brooks, G. Wolf, and Y. Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=CD9Snc73AW. Expert Certification.

H.-D. Tran, D. Manzanas Lopez, P. Musau, X. Yang, L. V. Nguyen, W. Xiang, and T. T. Johnson. Star-based reachability analysis of deep neural networks. In *International symposium on formal methods*, pages 670–686. Springer, 2019a.

H.-D. Tran, P. Musau, D. M. Lopez, X. Yang, L. V. Nguyen, W. Xiang, and T. T. Johnson. Parallelizable reachability analysis algorithms for feed-forward neural networks. In *2019 IEEE/ACM 7th International Conference on Formal Methods in Software Engineering (FormaliSE)*, pages 51–60. IEEE, 2019b.

H.-D. Tran, X. Yang, D. Manzanas Lopez, P. Musau, L. V. Nguyen, W. Xiang, S. Bak, and T. T. Johnson. NNV: the neural network verification tool for deep neural networks and learning-enabled cyber-physical systems. In *International conference on computer aided verification*, pages 3–17. Springer, 2020.

H.-D. Tran, S. W. Choi, Y. Li, H. Okamoto, B. Hoxha, and G. Fainekos. ProbStar Temporal Logic for Verifying Complex Behaviors of Learning-enabled Systems. In *Proceedings of the 28th ACM International Conference on Hybrid Systems: Computation and Control*, HSCC '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400715044. doi: 10.1145/3716863.3718035. URL https://doi.org/10.1145/3716863.3718035.

A. M. Tumlin, S. Sasaki, B. Wooding, D. M. Lopez, M. U. Zubair, N. Hashemi, H. Zhang, W. Abbas, I. Oguz, M. Ma, and T. T. Johnson. NNV3: Expanding Neural Network Verification to New Architectures and Domains. *arXiv preprint arXiv:2609.30050*, 2026.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*. Springer US, United States, 2005. ISBN 0387001522. doi: 10.1007/b106715.

F. Wang, P. Xu, W. Ruan, and X. Huang. Towards verifying the geometric robustness of large-scale neural networks. In *AAAI*, 2023.

S. Wang, H. Zhang, K. Xu, X. Lin, S. Jana, C.-J. Hsieh, and J. Z. Kolter. Beta-CROWN: Efficient bound propagation with per-neuron split constraints for complete and incomplete neural network verification. *Advances in Neural Information Processing Systems*, 34, 2021.

S. Webb, T. Rainforth, Y. W. Teh, and M. P. Kumar. A statistical approach to assessing neural network robustness. *arXiv preprint arXiv:1811.07209*, 2018.

K. Xu, H. Zhang, S. Wang, Y. Wang, S. Jana, X. Lin, and C.-J. Hsieh. Fast and Complete: Enabling complete neural network verification with rapid and massively parallel incomplete verifiers. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=nVZtXBI6LNn`.

G. Yang, T. Duan, J. E. Hu, H. Salman, I. Razenshteyn, and J. Li. Randomized smoothing of all shapes and sizes. In *International conference on machine learning*, pages 10693–10705. PMLR, 2020.

C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang. BiSeNet: Bilateral Segmentation Network for Real-Time Semantic Segmentation. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer Vision – ECCV 2018*, pages 334–349, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01261-8.

Y. Zhang, P. Kouvaros, and A. Lomuscio. Scalable Neural Network Geometric Robustness Validation via Hölder Optimisation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL `https://openreview.net/forum?id=Va1IHqY3jG`.

E. Zisselman and A. Tamar. Deep residual flow for out of distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13994–14003, 2020.

## A Proof of Theorem 1

We restate the setup. Let $\mathbf{R}_q := \{y \in \mathbb{R}^n : s(y;\, x) \le q\}$ be the probabilistic reach set obtained from the Hashemi double-step calibration of Section 3.2, with miscoverage level $\epsilon \in (0, 1)$, calibration size $m$, and rank $\ell \le m$. By the guarantee of Hashemi et al. [2025b], the conformal coverage event

$$\mathcal{E}_{\text{cov}} \;:=\; \Big\{ \Pr_{x\sim\mathcal{W}}[f(x) \notin \mathbf{R}_q] \le \epsilon \Big\} \tag{10}$$

holds with probability at least $\delta_2 = 1 - \mathsf{betacdf}_{1-\epsilon}(\ell, m+1-\ell)$ over the calibration sample. Let $\hat{\pi}_{\text{upper}}$ be the asymptotic upper bound on the in-ball unsafe mass $\pi := \Pr_{z\sim\pi_q}[\psi(z;\, x) \in \mathbf{U}]$ produced by bounded AMLS (Section 3.3). Under Assumption 2, the AMLS bound event

$$\mathcal{E}_{\text{AMLS}} \;:=\; \big\{ \pi \le \hat{\pi}_{\text{upper}} \big\} \tag{11}$$

holds with probability at least $1 - \delta_3$ over the AMLS sample. The event $\mathcal{E}_{\text{cov}}$ is a property of $q$ alone, which is determined by the calibration draw. Conditional on $q$, the AMLS chain runs on an independently seeded sample, so $\Pr[\mathcal{E}_{\text{AMLS}} \mid q] \ge 1 - \delta_3$ uniformly in $q$ under Assumption 2. The tower property then gives $\Pr[\mathcal{E}_{\text{cov}} \cap \mathcal{E}_{\text{AMLS}}] = \mathbb{E}_q[\mathbb{1}_{\mathcal{E}_{\text{cov}}}(q)\, \Pr[\mathcal{E}_{\text{AMLS}} \mid q]] \ge (1-\delta_3)\, \Pr[\mathcal{E}_{\text{cov}}] \ge \delta_2(1-\delta_3)$.

**Flow-fidelity assumption.** We assume the trained flow's conditional pushforward agrees with the true input-conditional output distribution restricted to $\mathbf{R}_q$. Concretely, with $p(\cdot \mid x)$ denoting the true distribution of $f(x)$ for $x \sim \mathcal{W}$ and $p_{\text{flow}}(\cdot \mid x)$ the flow's pushforward,

$$p(y \mid x) \;=\; p_{\text{flow}}(y \mid x) \quad \text{for almost every } y \in \mathbf{R}_q. \tag{12}$$

This assumption identifies the true-pushforward in-ball unsafe mass with $\pi$, which is the quantity bounded AMLS estimates.

**Proof.** Condition on $\mathcal{E}_{\text{cov}} \cap \mathcal{E}_{\text{AMLS}}$, which holds with probability at least $\delta_2(1-\delta_3)$. Decompose the unsafe-region probability under $x \sim \mathcal{W}$ over the partition $\{f(x) \in \mathbf{R}_q,\ f(x) \notin \mathbf{R}_q\}$:

$$\Pr_{x\sim\mathcal{W}}[f(x) \in \mathbf{U}] = \Pr[f(x) \in \mathbf{U},\ f(x) \in \mathbf{R}_q] + \Pr[f(x) \in \mathbf{U},\ f(x) \notin \mathbf{R}_q] \tag{13}$$

$$= \Pr[f(x) \in \mathbf{U} \mid f(x) \in \mathbf{R}_q] \cdot \Pr[f(x) \in \mathbf{R}_q] + \Pr[f(x) \in \mathbf{U},\ f(x) \notin \mathbf{R}_q]. \tag{14}$$

By (12), the conditional distribution of $f(x)$ given $f(x) \in \mathbf{R}_q$ equals $\pi_q$, so

$$\Pr[f(x) \in \mathbf{U} \mid f(x) \in \mathbf{R}_q] \;=\; \pi \;\le\; \hat{\pi}_{\text{upper}} \;\le\; \eta, \tag{15}$$

using the verdict-gate condition $\hat{\pi}_{\text{upper}} \le \eta$ and the AMLS bound $\pi \le \hat{\pi}_{\text{upper}}$ on $\mathcal{E}_{\text{AMLS}}$. The second term in (14) is bounded above by $\Pr[f(x) \notin \mathbf{R}_q] \le \epsilon$ on $\mathcal{E}_{\text{cov}}$. Substituting:

$$\Pr_{x\sim\mathcal{W}}[f(x) \in \mathbf{U}] \le \eta \cdot \Pr[f(x) \in \mathbf{R}_q] + \Pr[f(x) \notin \mathbf{R}_q] \tag{16}$$

$$= \eta \cdot \big(1 - \Pr[f(x) \notin \mathbf{R}_q]\big) + \Pr[f(x) \notin \mathbf{R}_q] \tag{17}$$

$$= \eta + (1-\eta) \cdot \Pr[f(x) \notin \mathbf{R}_q] \tag{18}$$

$$\le \eta + (1-\eta)\epsilon \tag{19}$$

$$= 1 - (1-\epsilon)(1-\eta) \tag{20}$$

$$\le \epsilon + \eta, \tag{21}$$

where the last inequality follows from $(1-\epsilon)(1-\eta) = 1-\epsilon-\eta+\epsilon\eta \ge 1-\epsilon-\eta$. This establishes (9) on $\mathcal{E}_{\text{cov}} \cap \mathcal{E}_{\text{AMLS}}$, which holds with probability at least $\delta_2(1-\delta_3)$ by the conditioning argument above. □

## B Method-section technical detail

**Conditional flow-matching loss.** Conditional flow matching [Lipman et al., 2023] replaces the marginal regression target with a closed-form per-sample conditional velocity $u_t(\cdot\,|y_1)$, giving the tractable loss

$$\mathcal{L}_{\text{CFM}}(\theta) \;=\; \mathbb{E}_{t,\,y_1,\,y_t\sim p_t(\cdot|y_1)} \big\| u_\theta(y_t, t) - u_t(y_t \mid y_1) \big\|^2, \tag{22}$$

which has the same gradient as the idealised flow-matching loss.

**Instantaneous change of variables.** The trained flow's exact log-density along the reverse trajectory is

$$\log p_1(y) \;=\; \log p_0(\phi(y)) - \int_0^1 \operatorname{div}_y u_\theta\big(\psi_t^{-1}(y),\, t\big)\, dt. \tag{23}$$

**Latent norm against the log-density score.** The flow's log-density is the natural alternative to the latent norm, and its super-level set at a given coverage is the minimum-volume region, so a comparison against it measures the latent norm against the theoretically optimal admissible score. Table 4 compares the two, together with calibrated fixed shapes, on the four proof-of-concept geometries whose ground-truth reach-set volume is known, at a coarser miscoverage level ($\alpha = 0.01$) than the paper operating point and with three seeds per cell. The two learned scores stay within a few percent of each other throughout, the latent norm ahead on the two rotated-linear geometries, level on banana-2D, and behind on three-blob-3D by less than the seed spread, at 6–19% lower wall-clock cost. Both learned scores separate from the fixed shapes exactly where the geometry turns curved or multimodal, and on banana-2D the $\ell_2$ ball is worse than the hyperrectangle, since no single convex body follows a curved region however it is oriented. Empirical coverage meets the $1-\alpha$ target on every cell.

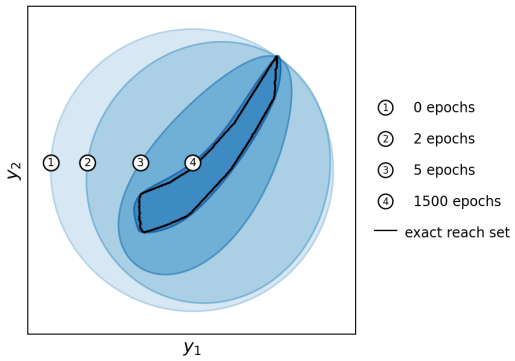



Figure 3: A visualization of the learned reach set from our method via flow matching. Through the training process, the reach set is refined to fit the geometry of the true reach set.

Table 4: Volume ratio $\mathrm{vol}(\mathbf{R}_{\text{predicted}})/\mathrm{vol}(\mathbf{R}_{\text{true}})$ (mean $\pm$ sd over 3 seeds, $\alpha = 0.01$) for the latent-norm score, the flow's log-density score, and calibrated fixed shapes on proof-of-concept geometries with known ground-truth volume. Smaller is tighter and 1.00 is exact.

| **Geometry** | **Latent norm** | **Log-density** | **Hyperrect.** | **$\ell_2$ ball** |
|---|---|---|---|---|
| rotated-linear-2D | $1.00 \pm 0.00$ | $1.07 \pm 0.00$ | $1.65 \pm 0.04$ | $1.38 \pm 0.03$ |
| rotated-linear-3D | $1.02 \pm 0.01$ | $1.11 \pm 0.00$ | $2.60 \pm 0.03$ | $1.81 \pm 0.02$ |
| banana-2D | $1.42 \pm 0.04$ | $1.42 \pm 0.02$ | $4.69 \pm 0.09$ | $4.99 \pm 0.08$ |
| three-blob-3D | $5.52 \pm 1.35$ | $5.41 \pm 1.54$ | $352 \pm 47$ | $349 \pm 50$ |

**Explicit integral form of the mass formulation.** The in-ball unsafe mass $\pi$ from (7) expands to

$$\pi \;=\; Z^{-1} \int_{\mathbf{B}_q} \mathbb{1}\{\psi(z;\, x) \in \mathbf{U}\}\, p_0(z)\, dz, \tag{24}$$

where $p_0 = \mathcal{N}(0, I_n)$ is the standard Gaussian on $\mathbb{R}^n$ and $Z = \Pr_{z\sim p_0}[\|z\|_2 \le q]$ is the probability mass that $p_0$ assigns to $\mathbf{B}_q$. The factor $Z^{-1}$ rescales the integral so that $\pi_q := p_0(\cdot \mid \|z\|_2 \le q)$ is a probability measure on $\mathbf{B}_q$ rather than $p_0$ restricted to that ball.

**Bounded-AMLS CLT transfer.** The bounded chain inherits the unbounded CLT of Cérou and Guyader [2016] only after re-verifying their regularity hypothesis (their Assumption [Ha]) for the in-ball-restricted Modified Metropolis kernel. We state the transfer as an explicit assumption.

**Assumption 2** (Bounded-AMLS asymptotic confidence interval)**.** The bounded AMLS estimator $\hat{\pi}$ admits a Cérou-Guyader-type asymptotic upper confidence bound $\hat{\pi}_{\text{upper}}(N, \delta_3)$, depending on the per-level particle count $N$ and confidence parameter $\delta_3$, such that $\Pr[\pi \le \hat{\pi}_{\text{upper}}] \to 1 - \delta_3$ as $N \to \infty$. Equivalently, the conclusion of Cérou and Guyader [2016] continues to hold for the in-ball-restricted Modified Metropolis kernel with the in-ball acceptance predicate.

A full derivation requires checking three conditions on the bounded kernel. The first is $\mu$-reversibility on $\mathbf{B}_q$, which is immediate because the in-ball predicate preserves detailed balance with respect to the truncated Gaussian $\pi_q$. The second is finite second moments of the level-set surface integrals intersected with $\mathbf{B}_q$, which is immediate because $\mathbf{B}_q$ is a bounded Borel set and the surface integrand is bounded on $\mathbf{B}_q$. The third is a local-equicontinuity-in-level $L^2$ regularity of the surface integrand restricted to the bounded chain. This third condition is non-trivial near the boundary $\partial\mathbf{B}_q$ where the indicator $\mathbb{1}_{\mathbf{B}_q}$ introduces a discontinuity, and a published derivation of the bounded-AMLS CLT remains absent. We leave the full derivation to future work.

**Worked example for the joint guarantee.** To make Theorem 1 concrete, fix conformal miscoverage $\epsilon = 10^{-3}$ and AMLS tail target $\eta = 10^{-4}$. The theorem certifies $\Pr_{x\sim\mathcal{W}}[f(x) \in \mathbf{U}] \le 1-(1-10^{-3})(1-10^{-4}) \approx 1.1\times10^{-3}$, essentially additive in this regime ($\epsilon+\eta = 1.1\times10^{-3}$, with multiplicative correction $-\epsilon\eta = -10^{-7}$ contributing negligibly). The corresponding joint confidence $\delta_2(1-\delta_3)$ depends on the calibration sample size and the confidence level chosen for the AMLS

Table 5: Error budget of the certificate by pipeline stage. The certified bound accumulates additively as $\epsilon + \eta + \tau$, one term per stage, and the two confidences compose multiplicatively as $\delta_2(1 - \delta_3)$. Calibration error is finite-sample and exact at any $m$, only the AMLS interval is asymptotic, and flow-training error affects tightness but never coverage validity.

| Stage | Error term | Confidence | Controlled by | Assumes |
|---|---|---|---|---|
| Flow training | $\tau$ (Theorem 4) | — | training budget | i.i.d. draws from $\mathcal{W}$ |
| Conformal calibration | $\epsilon$ | $\delta_2$ | $(m, \ell)$ | exchangeable, continuously distributed scores |
| Bounded AMLS | $\eta$ | $1 - \delta_3$ | particle budget | Assumption 2, one condition open |

bound. With $m = 8000$ and $\ell = 7999$ giving $\delta_2 \approx 0.997$, and the AMLS upper bound taken at the one-sided $1 - \delta_3 = 0.999$ level, the joint confidence is $\approx 0.996$. The conformal parameters $(\epsilon, m, \ell)$ control the in-coverage statement and the AMLS parameters $(\eta, \delta_3)$ control the in-coverage tail bound, so tightening either factor moves the bound or the confidence independently. Table 5 summarizes where each error term enters the certificate and which parameter controls it. The bound accumulates additively as $\epsilon + \eta + \tau$, one term per stage, while the two confidences compose multiplicatively as $\delta_2(1 - \delta_3)$.

# C Handling OR-of-AND specifications and disjoint input regions

Section 3 develops bounded AMLS for the case of a single connected unsafe region $\mathbf{U}$ over a single connected input region $\mathbf{I}$. Two practical extensions arise on standard benchmarks. The first is unsafe specifications that are unions of $K$ halfspace conjunctions (the OR-of-AND format used by VNN-LIB), and the second is input regions partitioned into multiple disjoint boxes (e.g., ACAS Xu prop_3 with four disjoint input regions). We describe the aggregation and verdict logic for both. Both extensions preserve Theorem 1 with the modifications stated below.

## C.1 Union AMLS for $K$-disjunct OR specifications

Suppose the unsafe region is a union of $K$ halfspace conjunctions, $\mathbf{U} = \bigcup_{j=1}^{K} \mathbf{U}_j$. The naive approach is to run $K$ independent bounded-AMLS chains, one per disjunct $\mathbf{U}_j$, and combine their results via a Bonferroni union bound. If each chain certifies the per-disjunct in-ball mass $\pi_j \leq \eta$ with confidence $1 - \delta_3$, then $\pi := \Pr_{z \sim \pi_q}[\psi(z; x) \in \mathbf{U}] \leq \sum_j \pi_j \leq K\eta$ with confidence $1 - K\delta_3$. The Bonferroni penalty wastes a factor of $K$ in tightness on the joint guarantee and requires $K$ separate flow-pushforward evaluations per AMLS step.

The *union variant* of bounded AMLS replaces the per-disjunct margin function $\varphi_j$ with the elementwise minimum

$$\varphi_{\text{union}}(y) \;:=\; \min_{j=1,\ldots,K} \varphi_j(y), \tag{25}$$

which satisfies $\varphi_{\text{union}}(y) \leq 0 \Leftrightarrow y \in \mathbf{U}$ by construction. A single AMLS chain instantiated with $\varphi_{\text{union}}$ then estimates $\pi_{\text{union}} := \Pr_{z \sim \pi_q}[\psi(z; x) \in \mathbf{U}]$ directly, bypassing the Bonferroni step. The verdict-gate logic and the joint guarantee of Theorem 1 apply unchanged with $\pi$ replaced by $\pi_{\text{union}}$. The cost is one min-reduction per pushforward evaluation, which is negligible compared to the flow ODE solve.

**Validity of the union variant.** The Au-Beck Modified Metropolis Algorithm and the Cérou-Guyader CLT both require only that the level-set indicators $\{z : \varphi(\psi(z; x)) \leq \tau_k\}$ be measurable and have positive measure under $\pi_q$. Both conditions hold for $\varphi_{\text{union}}$: the elementwise minimum of finitely many measurable margin functions is measurable, and the level set $\{y : \varphi_{\text{union}}(y) \leq \tau_k\}$ equals the union $\bigcup_j \{y : \varphi_j(y) \leq \tau_k\}$, which has positive measure whenever any single disjunct does. The level decomposition therefore goes through, and the asymptotic confidence interval applies to $\pi_{\text{union}}$ directly.

**When to use which variant.** We use the union variant for benchmarks with high-cardinality OR specifications, namely metaroom_2023 ($K = 19$), vit_2023 ($K = 9$), tinyimagenet_2024 ($K = 199$),

and cifar100_2024 ($K = 99$). For benchmarks with $K \leq 2$ we run the per-disjunct variant. The cost is comparable, and the per-disjunct $\hat{\pi}_{\text{upper}}$ values are useful diagnostics for understanding which class boundary contributes most to the verdict.

### C.2 Per-box runs and union-bound aggregation for disjoint input regions

When the input region is partitioned into $B$ disjoint boxes $\mathbf{I} = \mathbf{I}_1 \sqcup \cdots \sqcup \mathbf{I}_B$ (e.g., ACAS Xu prop_3 has $B = 4$ disjoint input regions), we treat each box as an independent verification subproblem. For each $b = 1, \ldots, B$ we train one flow $\psi_t^{(b)}$ on $\mathbf{I}_b$, calibrate $q^{(b)}$, and run bounded AMLS with per-box parameters $(\epsilon^{(b)}, m^{(b)}, \ell^{(b)}, \eta^{(b)}, N_{\text{AMLS}}^{(b)})$, yielding $\delta_2^{(b)}$ and $\delta_3^{(b)}$ via the Hashemi double-step and Assumption 2 respectively, to obtain a per-box verdict $v_b \in \{\text{safe}, \text{unknown}\}$.

The aggregate verdict over the partition is determined by the per-box verdicts as follows. The pipeline returns *safe* on the input partition iff every per-box pipeline returns *safe*. With the input distribution $\mathcal{W}$ decomposed as $\mathcal{W} = \sum_{b=1}^{B} w_b \mathcal{W}^{(b)}$ for $\mathcal{W}^{(b)}$ supported on $\mathbf{I}_b$ and weights $w_b = \Pr_{x \sim \mathcal{W}}[x \in \mathbf{I}_b]$, the aggregate guarantee is

$$\Pr_{x \sim \mathcal{W}}[f(x) \in \mathbf{U}] \;=\; \sum_{b=1}^{B} w_b \cdot \Pr_{x \sim \mathcal{W}^{(b)}}[f(x) \in \mathbf{U}] \;\leq\; \sum_{b=1}^{B} w_b \left(\epsilon^{(b)} + \eta^{(b)}\right), \tag{26}$$

holding with confidence at least $\prod_{b=1}^{B} \delta_2^{(b)}(1 - \delta_3^{(b)})$ (the per-box calibration and AMLS confidence events are independent by construction). At our default operating point with $\delta_2^{(b)}(1 - \delta_3^{(b)}) \approx 0.996$ for every box, ACAS Xu prop_3 with $B = 4$ gives an aggregate confidence of $0.996^4 \approx 0.984$, which is the right number to compare against the empirical false-UNSAT rates in Appendix H for prop_3 instances. The pipeline returns *unknown* when any per-box verdict is unknown.

When the experimental harness pairs the pipeline with a falsifier (Section 4), an *unsafe* verdict produced by the falsifier on any single box short-circuits the partition's verdict to unsafe with the produced counterexample. The flow lane alone never returns *unsafe*, so this case arises only via the experimental falsifier coupling and is not a methodological contribution.

## D Evaluation details

This appendix records replication detail displaced from Section 4.1.

**Bounded-AMLS knob defaults.** The bounded-AMLS specification check has four operational knobs introduced in Section 3.3: the tail target $\eta$, the target conditional level rate $\rho$, the number of MCMC steps per level, and the MCMC step size with an optional $q/\sqrt{n}$ adaptive variant. We use $\eta = 10^{-4}$, $\rho = 0.5$, 50 MCMC steps per level, and the $q/\sqrt{n}$ adaptive step size as the default configuration across benchmarks.

**Per-benchmark training and AMLS budgets.** A practical benefit of the pipeline is that the flow-training and AMLS budgets can be tuned per benchmark to match available compute. Benchmarks with larger output dimension or tighter VNN-COMP timeouts use smaller $n_{\text{train}}$ or fewer epochs to fit within the per-benchmark budget. Concrete per-benchmark settings are reported in Table 6.

**Empirical-coverage formula.** We compute empirical coverage of $\mathbf{R}_q$ on a held-out test sample,

$$\widehat{\text{cov}}_{\text{emp}} \;:=\; \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \mathbb{1}\{f(x_i^{\text{test}}) \in \mathbf{R}_q\}, \qquad x_i^{\text{test}} \overset{\mathcal{W}}{\sim} \mathbf{I}, \tag{27}$$

with the test sample drawn i.i.d. from $\mathcal{W}$ on $\mathbf{I}$ and disjoint from training and calibration. The empirical-coverage check compares $\widehat{\text{cov}}_{\text{emp}}$ to the theoretical target $1 - \epsilon$ at every (benchmark, method) cell as a direct empirical validation of conformal correctness.

**Joint-guarantee consistency check.** The joint-guarantee check compares the empirical false-UNSAT rate in each (method, benchmark) cell against the per-instance confidence budget $1 - \delta_2(1 - \delta_3) \approx 0.004$ and the certificate's probabilistic semantics, with results in Appendix H.

Table 6: Per-benchmark training and verification budgets for our pipeline. All cells share $\epsilon = 10^{-3}$, $\eta = 10^{-4}$, $m = 8000$, $\ell = 7999$, AMLS max levels $= 30$, and target conditional rate $\rho = 0.5$. Benchmarks with $K$-disjunct robustness specifications (`metaroom_2023`, `vit_2023`, `tinyimagenet_2024`, `cifar100_2024`) use the union variant of bounded AMLS that folds the $K$-disjunct OR into a single chain (Appendix C).

| **Benchmark** | $n_{\text{train}}$ | **Flow epochs** | **Scenario** $n$ | **Per-instance timeout** |
|---|---|---|---|---|
| `acasxu_2023` | 5,000 | 2,000 | 2,000 | 116 s |
| `collins_rul_cnn_2022` | 1,000 | 1,000 | 500 | 30–1,800 s (per instance) |
| `dist_shift_2023` | 10,000 | 2,000 | 2,000 | 300 s |
| `linearizenn_2024` | 10,000 | 2,000 | 2,000 | 900 s |
| `tllverify_2023` | 10,000 | 2,000 | 2,000 | 600 s |
| `metaroom_2023` | 10,000 | 2,000 | 2,000 | 210 s |
| `vit_2023` | 10,000 | 2,000 | 2,000 | 100 s |
| `tinyimagenet_2024` | 10,000 | 2,000 | 2,000 | 100 s |
| `cifar100_2024` | 10,000 | 2,000 | 2,000 | 100 s |

Table 7: Volume comparison on the synthetic benchmarks. Each cell reports the mean volume ratio over 5 seeds, where ratio $= \text{vol}(R_{\text{predicted}})/\text{vol}(R_{\text{true}})$. Smaller is tighter, and $1.00$ is exact. The reach-set ground truth comes from either a closed-form determinant (synth_$N$d, identity-activation 1-Lipschitz nets) or a cached $N{=}10^7$-sample MC estimate (the bananas). Sample-budget configs are small/medium/large = $m{=}1\text{k}/8\text{k}/16\text{k}$. *Star (approx)* is the sound star-set over-approximation baseline and has no $m$ axis.

| | **Ours** | | | **Clipping-Block** | | | **Star** |
|---|---|---|---|---|---|---|---|
| **Benchmark** | small | medium | large | small | medium | large | approx |
| 2d_banana | 4.72 | 1.33 | 1.09 | 4.39 | 4.37 | 4.36 | 5.19 |
| synth_2d | 6.90 | 1.31 | 1.08 | 6.88 | 6.90 | 6.95 | 7.00 |
| 3d_banana | 345 | 9.37 | 4.27 | 413 | 418 | 425 | 840 |
| synth_3d | 533 | 11.81 | 3.32 | 552 | 606 | 588 | 590 |
| synth_5d | 762 | 20.88 | 6.12 | 2.4 e3 | 2.3 e3 | 2.4 e3 | 2.9 e3 |
| synth_10d | 7.1 e5 | 434 | 161 | 3.0 e7 | 3.5 e7 | 4.7 e7 | 9.8 e7 |
| synth_20d | 1.1 e8 | 1.2 e5 | 1.5 e5 | 1.4 e12 | 5.3 e12 | 8.3 e12 | 3.7 e15 |

**Design rationale for the 1-Lipschitz scaling family.** The networks are ReLU MLPs whose linear layers are spectrally normalized to operator norm at most 1, so the composition is 1-Lipschitz overall. Pairing this network with a halfspace specification beyond the image of $\mathbf{I}$ gives an UNSAT-by-construction family in which the ground-truth verdict is fixed by the Lipschitz constraint, and the headline question reduces to verification cost as a function of depth alone.

**Synthetic volume comparison.** On a family of low-dimensional synthetic benchmarks where the exact reach-set volume is computable (closed-form determinant for the identity-activation 1-Lipschitz nets, cached high-sample Monte Carlo estimate for the bananas), Table 7 reports the volume ratio $\text{vol}(\mathbf{R}_{\text{predicted}})/\text{vol}(\mathbf{R}_{\text{true}})$ for our flow-based reach set, the Clipping-Block baseline, and the sound star-set over-approximation. Smaller is tighter and $1.00$ is exact. The flow-based reach set tightens monotonically with calibration sample budget, while the fixed-shape baselines plateau at the geometry mismatch their score induces.

# E Protocol and audits for certification beyond convex-friendly perturbation models

This appendix records the protocol behind Table 2 and the audits behind its claims.

**ResNet-50 composed transformation model.** The input set composes four geometric parameters (rotation $\pm 5^\circ$, scale $\pm 2\%$, translation $\pm 2\%$ along each axis) with seven photometric ones (brightness $\pm 0.02$, contrast $\pm 3\%$, gamma $\pm 3\%$, saturation $\pm 5\%$, per-channel RGB gains $\pm 2\%$), all eleven varying jointly and uniformly over the box. The verified map sends a parameter vector $\theta \in \mathbb{R}^{11}$

to the ResNet-50 logits of the correspondingly transformed image. The 50 images are taken as an even spread across the clean top-5-margin range (0.98–24.3) of the first 200 images of a fixed seed-permuted manifest of correctly classified ImageNet validation images. A falsifier (4,000 random draws plus PGD over $\theta$) runs on every instance for both methods, and an instance it does not falsify counts only as having no known counterexample. Clipping-Block runs as proposed, a PCA-reduced sampled convex hull with per-coordinate conformal inflation at $m = 8000$ and confidence $0.997$, while ours runs at the paper operating point (joint confidence $0.996$). A pre-registered abstention rule declines to certify when the chain's in-ball margin $\varphi$, the smallest specification margin the AMLS search observes inside the conformal ball, falls below $0.15$. The rule never triggered (minimum observed $\varphi = 0.219$). The contrast with the $\ell_\infty$ setting discussed in Section 4.2 is measurable in the calibration itself. The per-coordinate conformal inflation Clipping-Block requires on this network has median $0.10$ logits under small $\ell_\infty$ balls against $1.16$ logits under the composed transformation model, and on an adversarially trained $\ell_\infty$-robust ResNet-50 (200 instances spanning $\varepsilon = 1/255$ to $8/255$) the two methods' verdicts agree on 200 of 200 instances.

**ResNet-50 audit.** Every chain certificate was audited at $50{,}000$ fresh draws plus per-coordinate PGD, covering 71 certificates (43 at top-5, 28 at top-1) and $3{,}550{,}000$ network evaluations in total. Of the 71 certificates, 68 showed zero violating draws, and a Bonferroni-corrected one-sided 95% upper bound across all 71 is $1.5 \times 10^{-4}$, below the certified $1.1 \times 10^{-3}$. Two certificates carried violating mass below the bound ($2 \times 10^{-5}$ and $4 \times 10^{-4}$), consistent with their certificates. One exceeded it, at measured mass $1.4 \times 10^{-3}$ with 95% lower bound $1.09 \times 10^{-3}$, and we report it as an empirical certificate failure. Both chain certificates issued on falsified instances (one per specification, counted in Table 2's last column) are reported as SAT under falsifier pairing. On the instances the falsifier broke, the chain independently declined on 18 of 19 at top-1 and 7 of 8 at top-5.

**BiSeNet segmentation protocol.** We verify the `BiSeNet.onnx` CamVid model released with the Clipping-Block paper (12 classes at $720 \times 960$, output dimension 8.3M) for per-pixel argmax preservation under the photometric sub-family at twice the ResNet-50 half-widths. Pixels are drawn from an interior lattice of up to 96 points per image at least 4 px from the clean segmentation boundary, fixed as a function of the clean image only, giving 2,136 audited pixels across 25 images. Of these, 1,296 survive the attack lane of $20{,}000$ uniform draws plus PGD per pixel, with the same draws serving as each issued certificate's audit sample. The abstention threshold is a hyperparameter of this experiment's protocol rather than a component of the method, and we report how it was set because the certified counts depend on it. The gate $\varphi \geq 3$ was selected by Learn-then-Test [Angelopoulos et al., 2025] fixed-sequence testing over the grid $t \in \{6, 4, 3, 2, 1.5, 1, 0.5, 0.15\}$ at target level $0.02$, using the Clopper-Pearson 95% upper bound on $\Pr[\text{attack-SAT} \mid \text{certified},\ \varphi \geq t]$ computed on 12 held-out CamVid images disjoint from the evaluation set, and applied frozen.

**BiSeNet baseline representations.** Clipping-Block's released pipeline calibrates on the full 8.3M-dimensional output field, and on the audited pixels it certifies 408 of 1,296 (31.5%). Table 2 instead grants it the specification representation, calibrating on the logits the per-pixel specification queries, which raises it to 762 (58.8%), and we report the stronger variant as the baseline. Without the abstention gate our pipeline certifies 1,283 of 1,296 audited pixels (99.0%), its certified set strictly contains both Clipping-Block variants' on every audited pixel, and it issues 66 certificates on pixels where the attack lane finds a counterexample. Those 66 are possible because the certificate and the attack answer different questions. The certificate bounds the probability that a random perturbation changes the pixel's label, PGD searches for any single perturbation that does, and a counterexample can sit in a region whose probability lies below the certified $1.1 \times 10^{-3}$ bound, so a certificate on an attacked pixel is not automatically invalid. It is still a certificate on a pixel with a known counterexample. The abstention gate extends the pipeline to prevent exactly these false UNSATs, withholding a certificate whenever the chain's in-ball margin falls below a threshold, and it applies to any task where held-out data is available to set the threshold. Here the threshold was selected by Learn-then-Test on 12 held-out CamVid images disjoint from the evaluation set and applied frozen, as described in the protocol above. The gate reduces the 66 false UNSATs to the 5 reported in Table 2, at the cost of 242 certificates on the attack-surviving pool.

**Sound-verifier boundary.** We ran $\alpha$,$\beta$-CROWN on both models on our hardware (A30, 24 GB). The native BiSeNet ONNX graph uses operators unsupported by the verifier's front end, so we built a supported-operator re-expression and verified it equivalent in-process (identical argmax at all 691,200

pixels, maximum logit deviation $1.5 \times 10^{-5}$). Verification runs out of GPU memory, matching the out-of-memory result reported for $\alpha$,$\beta$-CROWN on this model in the Clipping-Block paper at 48 GB. For ResNet-50 we encoded the photometric parameter space as an exact input layer composed with the network, with equivalence gates against the original model, and verification likewise runs out of memory.

# F TV-relaxed joint guarantee

The exact flow-fidelity assumption in Appendix A (12) is the conservative form of an underlying conditional-distribution agreement. We record here a quantitative relaxation that holds under the more realistic hypothesis of bounded conditional total-variation distance.

**Definition 3** (Conditional TV budget)**.** For a measurable subset $\mathbf{R}_q \subseteq \mathbb{R}^n$ with $\Pr_{x\sim\mathcal{W}}[f(x) \in \mathbf{R}_q] > 0$ and $\Pr_{y\sim p_{\text{flow}}}[y \in \mathbf{R}_q] > 0$, the trained flow has *conditional TV slack at most* $\tau \in [0,1]$ on $\mathbf{R}_q$ if

$$\sup_{A \text{ measurable}} \left| \Pr[f(x) \in A \mid f(x) \in \mathbf{R}_q] - \Pr_{y\sim p_{\text{flow}}}[y \in A \mid y \in \mathbf{R}_q] \right| \;\le\; \tau. \tag{28}$$

**Theorem 4** (TV-relaxed joint coverage + tail bound)**.** Under the same calibration and bounded-AMLS hypotheses as Theorem 1, but with the exact flow-fidelity assumption (12) replaced by the conditional TV budget (28) for some $\tau \ge 0$ with $\eta + \tau \le 1$, with confidence at least $\delta_2(1-\delta_3)$ over the joint randomness of calibration and AMLS,

$$\Pr_{x\sim\mathcal{W}}[f(x) \in \mathbf{U}] \;\le\; 1 - (1-\epsilon)(1-(\eta+\tau)) \;\le\; \epsilon + \eta + \tau. \tag{29}$$

*Proof sketch.* The conditional TV budget (28) gives $\Pr[f(x) \in \mathbf{U} \mid f(x) \in \mathbf{R}_q] \le \Pr_{y\sim p_{\text{flow}}}[y \in \mathbf{U} \mid y \in \mathbf{R}_q] + \tau \le \pi + \tau \le \eta + \tau$ on the AMLS bound event. Substituting $\eta + \tau$ for $\eta$ in the proof of Theorem 1 (Appendix A) yields (29). Setting $\tau = 0$ recovers Theorem 1. □

# G Full VNN-COMP'25 verdict matrix

Table 8 reports the full verdict breakdown ($\#\,\text{Verified} \,/\, \#\,\text{Falsified}$) alongside the % Solved metric for every (method, benchmark) cell. The compact UNSAT-recall view in Section 4 (Table 1) is a projection of this table. We include the full matrix here for readers who want SAT/UNSAT counts and the precise sound-violation tally per cell.

# H Empirical validation of the joint guarantee

This appendix reports the two empirical diagnostics flagged in Appendix D. The first measures how often the per-instance UNSAT verdict turns out to be wrong against the VNN-COMP'25 ground-truth consensus, and contrasts that empirical rate with the theoretical bound from Theorem 1. The second measures how often a held-out test output falls inside the calibrated reach set $\mathbf{R}_q$, and contrasts that mean with the theoretical coverage target $1-\epsilon$.

**Empirical false-UNSAT rate.** Table 9 reports, for each (method, benchmark) cell, the rate at which a UNSAT verdict turned out to be a false UNSAT against the ground-truth consensus. Theorem 1 permits such disagreements through two channels of different standing. A failure of the certificate's confidence event has probability at most $1-\delta_2(1-\delta_3) \approx 0.004$ per decided instance at our operating point. An instance whose unsafe region is non-empty but carries $\mathcal{W}$-measure below the certified $1.1 \times 10^{-3}$ bound, the icicle geometry discussed in Section 4.2, is instead reported UNSAT correctly under the probabilistic semantics the certificate speaks and counts as a false UNSAT only against the existential ground truth. The reported rates, highest at metaroom_2023 and tinyimagenet_2024 under $2.5\%$, exceed the $0.4\%$ confidence budget alone, and the remainder corresponds to these icicle instances, semantic disagreements rather than certificate failures.

**Empirical conformal coverage.** Table 10 reports the per-instance mean of the held-out coverage estimator $\widehat{\text{cov}}_{\text{emp}}$ defined in Appendix D. The theoretical target is $1-\epsilon = 0.999$. Our pipeline meets the target on every benchmark except metaroom_2023, where it averages $0.9707$, while Clipping-Block meets it everywhere, averaging $1.000$ on metaroom_2023. The metaroom_2023 deficit traces

Table 8: Verification results on VNN-COMP'25 benchmarks. Each cell reports % *Solved*↑ over the breakdown ($\#\,\mathrm{Verified}$ / $\#\,\mathrm{Falsified}$). % Solved is the fraction of instances *correctly* decided (verified UNSAT or falsified SAT, in agreement with the VNN-COMP'25 ground-truth consensus). A dagger ($\dagger N$) marks cells with $N$ verdicts that disagree with the consensus (sound violation). For our pipeline, $\#\,\mathrm{Verified}$ UNSATs are produced by the flow lane and $\#\,\mathrm{Falsified}$ SATs by the paired falsifier (PGD/APGD on $f$). Falsifier-produced SATs are concrete counterexamples on $f$ and thus always sound, so daggers in our row count only flow-lane false UNSATs.

| **Method** | ACAS Xu | collins rul-cnn | dist shift | linear-izenn | tll-verify | meta-room | vit | tiny imagenet | cifar 100 |
|---|---|---|---|---|---|---|---|---|---|
| **Ours** | 77.4%$^{\dagger 1}$ (136 / 9) | 100.0% (39 / 23) | 97.2% (63 / 7) | 91.7% (54 / 1) | 96.9% (15 / 16) | 98.0%$^{\dagger 2}$ (97 / 3) | 100.0% (200 / 0) | 98.0%$^{\dagger 4}$ (166 / 34) | 98.5%$^{\dagger 3}$ (173 / 26) |
| *Sound verifiers (from VNN-COMP'25)* | | | | | | | | | |
| $\alpha\beta$-CROWN | 100.0% (140 / 47) | 100.0% (40 / 23) | 100.0% (66 / 7) | 100.0% (60 / 1) | 100.0% (16 / 17) | 99.0% (95 / 5) | 62.7% (126 / 0) | 87.6% (138 / 38) | 64.7% (101 / 29) |
| NeuralSAT | 100.0% (140 / 47) | 100.0% (40 / 23) | 100.0% (66 / 7) | 100.0% (60 / 1) | 100.0% (16 / 17) | 99.0% (95 / 5) | 33.8% (68 / 0) | 77.1% (117 / 38) | 58.2% (88 / 29) |
| PyRAT | 99.5% (140 / 46) | 100.0% (40 / 23) | 98.6% (65 / 7) | 100.0% (60 / 1) | 100.0% (16 / 17) | 98.0%$^{\dagger 2}$ (98 / 3) | 41.8% (84 / 0) | 51.7% (69 / 35) | 43.3% (62 / 25) |
| CORA | 98.4% (138 / 46) | 100.0% (40 / 23) | 100.0% (66 / 7) | 100.0% (60 / 1) | 100.0% (16 / 17) | 97.0% (93 / 5) | – | 3.0% (1 / 5) | 5.5% (1 / 10) |
| *Probabilistic verifiers* | | | | | | | | | |
| Clipping-Block | 67.2% (116 / 9) | 98.4% (38 / 23) | 91.7% (59 / 7) | 100.0% (59 / 1) | 96.9% (15 / 16) | 92.0% (89 / 3) | 100.0% (200 / 0) | 98.0%$^{\dagger 4}$ (166 / 34) | 98.5%$^{\dagger 3}$ (174 / 26) |
| SAVER | 77.4%$^{\dagger 41}$ (179 / 6) | 100.0% (39 / 23) | 94.4%$^{\dagger 4}$ (69 / 3) | 98.3%$^{\dagger 1}$ (60 / 0) | 93.8%$^{\dagger 2}$ (17 / 15) | 2.0% (0 / 2) | 0.0% (0 / 0) | 0.5% (0 / 1) | 0.0% (0 / 0) |

Table 9: Empirical false-UNSAT rate per (method, benchmark) cell, reported over the breakdown ($\#\,\mathrm{false\,UNSAT}$ / $\#\,\mathrm{UNSAT\,verdicts}$). A false UNSAT is an UNSAT verdict on an instance whose VNN-COMP'25 ground-truth consensus is SAT. The consensus records the existence of a counterexample anywhere in the input region while the certificate bounds its measure, so a false UNSAT can coincide with a valid certificate.

| **Method** | ACAS Xu | collins rul-cnn | dist shift | linear-izenn | tll-verify | meta-room | vit | tiny imagenet | cifar 100 |
|---|---|---|---|---|---|---|---|---|---|
| **Ours** | 0.74% (1 / 136) | 0.00% (0 / 39) | 0.00% (0 / 63) | 0.00% (0 / 54) | 0.00% (0 / 15) | 2.06% (2 / 97) | 0.00% (0 / 200) | 2.41% (4 / 166) | 1.73% (3 / 173) |
| Clipping-Block | 0.00% (0 / 116) | 0.00% (0 / 38) | 0.00% (0 / 59) | 0.00% (0 / 59) | 0.00% (0 / 15) | 0.00% (0 / 89) | 0.00% (0 / 200) | 2.41% (4 / 166) | 1.72% (3 / 174) |
| SAVER | 22.91% (41 / 179) | 0.00% (0 / 39) | 5.80% (4 / 69) | 1.67% (1 / 60) | 11.76% (2 / 17) | – (0/0) | – (0/0) | – (0/0) | – (0/0) |

to the extreme-tiny-perturbation instances ($\varepsilon \in [4.36 \times 10^{-6},\, 10^{-5}]$) on which our flow's whitened output distribution concentrates so tightly that calibration overshoots and the held-out test sample lands marginally outside $\mathbf{R}_q$. Clipping-Block avoids this failure mode because its axis-aligned hyperrectangle is wider in each output coordinate.

Table 10: Per-cell empirical conformal coverage on a held-out test sample. Each cell reports the per-instance mean of $\widehat{\mathrm{cov}}_{\mathrm{emp}}$ over the breakdown $(\mathrm{mean}, n)$, where $n$ is the number of instances contributing to the mean. The theoretical target is $1 - \epsilon = 0.999$ for both methods. Instances whose held-out test sample yielded $\widehat{\mathrm{cov}}_{\mathrm{emp}} = 0$ are excluded as instrumentation artifacts (test sample concentrated outside $\mathbf{R}_q$ under whitening, observed on metaroom_2023's tiny-perturbation instances at $\varepsilon \leq 10^{-5}$). Cells that do not run conformal calibration (sound verifiers, SAVER, ProbStar, randomized smoothing) emit em-dash.

| **Method** | ACAS Xu | collins rul-cnn | dist shift | linear-izenn | tll-verify | meta-room | vit | tiny imagenet | cifar 100 |
|---|---|---|---|---|---|---|---|---|---|
| **Ours** | 0.9997 (n=177) | 0.9998 (n=39) | 0.9999 (n=65) | 0.9998 (n=59) | 0.9998 (n=16) | 0.9707 (n=71) | 0.9998 (n=200) | 0.9997 (n=166) | 0.9997 (n=173) |
| Clipping-Block | 0.9999 (n=177) | 0.9999 (n=38) | 0.9997 (n=65) | 0.9998 (n=59) | 0.9999 (n=16) | 1.0000 (n=97) | 0.9990 (n=200) | 0.9990 (n=166) | 0.9990 (n=174) |

## 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: Our first claim that our method results in tighter reach sets than existing probabilistic methods is supported by Figure 2b, which shows that our method produces reach-sets that are tighter (by volume) than fixed-shape baselines on synthetic benchmarks where the ground truth is computable, and by Table 1, which shows our method matches or exceeds the fixed-shape probabilistic baselines on most VNN-COMP benchmarks. Our second claim that our method scales better than sound verifiers is supported by Figure 2a, which directly shows how two of the top sound verifiers ($\alpha$,$\beta$-CROWN and NeuralSAT) fail to scale as the size of the neural network increases, and by Table 1, which shows that our method succeeds in verifying instances at VNN-COMP that no sound tool was able to. Our claim that the learned geometry certifies more than a calibrated convex baseline under composed geometric and photometric perturbations is supported by Table 2.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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: We conclude the work (Section 6) with a paragraph dedicated to describing the limitations of our approach, as well as future work that could be done to improve upon it.

   Guidelines:

   - The answer [N/A] 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: The paper has two theorems and one assumption. Theorem 1 is proved in Section A, with all hypotheses (calibration confidence, AMLS upper bound, flow fidelity) stated in the theorem block. The TV-relaxed companion Theorem 4 is stated and proved in Section F under the conditional total-variation budget defined there. Theorem 2 states the bounded-AMLS asymptotic confidence interval, with the conditions and limitations of the transfer from Cérou and Guyader [2016] discussed in Section B.

   Guidelines:

   - The answer [N/A] 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: We provide a thorough description of our approach Section 3, the preliminary information to understand it Section 2, and a visualization of the entire method Figure 1. In addition, the source code is publicly available at the link in the abstract, including instructions for reproducing the results.

   Guidelines:

   - The answer [N/A] 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 code, as well as instructions for reproducing the main experimental results, is publicly available at `https://github.com/sammsaski/n2v-neurips2026`.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: Experimental setup and choice of hyperparameters for our experiments are described in Section 4.1 and Section D

   Guidelines:

   - The answer [N/A] 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: [No]

   Justification: Each cell of the volume comparison (Figure 2b) is the mean volume ratio over 5 random network initializations of the same architecture family. Because the randomly-initialized 1-Lipschitz networks have widely varying reach-sets, the seed-to-seed spread reflects network-to-network volume variation rather than measurement noise of our pipeline, and reporting it as an error bar would not be informative about algorithm performance. The VNN-COMP UNSAT accuracy numbers in Table 1 aggregate over the full per-benchmark instance set (33–200 instances per benchmark depending on the benchmark) under each instance's published timeout, with statistical reliability coming from sample size. The joint safety certificate of Theorem 1 is itself a confidence statement at level $\delta_2(1-\delta_3)$ over the calibration sample and the AMLS particle budget.

   Guidelines:

   - The answer [N/A] 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: Section 4 describes the compute that was used to run the experiments.

Guidelines:

- The answer [N/A] 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 works for the analysis of neural networks and does not describe an approach that could be used in any way to damage, inconvenience, or infringe upon others. Additionally, it does not contain any private information about others, and all code will be made publicly available.

Guidelines:

- The answer [N/A] 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: The main motivation of this line of work (neural network verification) is for the safe deployment of neural networks in safety-critical systems, particularly because of their increased use within them. The approach in this paper improves on existing neural network verification techniques, enabling better safety analysis of them, and acknowledges this early in the paper and in the discussion of the results.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper doesn't involve release of data or models that have high risk for misuse.

    Guidelines:

    - The answer [N/A] 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: All creators or original owners of assets used in our experiments are properly credited throughout. Section 4.1 explicitly describes the benchmarks used for the experiments.

    Guidelines:

    - The answer [N/A] 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 only new asset introduced in the paper is the code, which has documentation on how to run and reproduce the results in the paper.

Guidelines:

- The answer [N/A] 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: [N/A]

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

Guidelines:

- The answer [N/A] 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: [N/A]

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

Guidelines:

- The answer [N/A] 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 rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not involved in the core method development for this research.

Guidelines:

- The answer [N/A] 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 in the NeurIPS handbook for what should or should not be described.