Energy-Shaped Manifold Projections Enable Adversarial Detection

Anonymous Author(s)
Affiliation
Address
email

Abstract

Adversarial attacks and distribution shift undermine reliability of deep classifiers. We revisit energy-based out-of-distribution (OOD) detection and propose a simple projection head that maps representations onto a learned data manifold and uses the squared norm of the projected vector as an energy score. The training is parallel with classification loss on the classification head and soft energy separation loss on the projection head that pushes adversarial examples to high energy while keeping clean examples at low energy. On a CIFAR-10 (Krizhevsky [2009]) variant with a held-out 10th class acting as OOD, our method detects both fast gradient sign (FGSM) and projected gradient descent (PGD) adversarial examples even when the classifier remains non-robust. We study design choices, including hinge versus softplus energy losses, regularization on the projected vector and the importance of normalization layer choice to align train and test statistics. Despite energy separation transferring across attacks, we find little OOD rejection of unrelated images and highlight failure modes. Our work provides a critical analysis of energy-shaped projections and lays out open problems and possibilities for future research.

1 Introduction

Machine learning systems deployed in high-stakes applications must cope with unreliable data: inputs may be perturbed by adversaries, drawn from shifted distributions, contain missing or biased values, or arise from human interaction. Standard training objectives optimize for accuracy, but offer no guarantees when inputs deviate from the training distribution. Recent work emphasizes OOD input detection as a complementary strategy to robust classification. Energy-based scores, derived from the log partition function, have been shown to distinguish OOD samples (Liu et al. [2020]).

In this paper, we revisit energy-based detection for adversarial perturbations and present the energy-shaped manifold projection head. The method maps the last hidden representation z from a standard backbone to a lower- or the same-dimensional representation z'; the squared norm $E = \|z'\|^2$ is used as an energy score. The soft separation loss encourages low energy for clean examples and high energy for adversarial data while classification is trained in parallel. We implement flexible loss functions (ReLU (hinge), softplus, and squared hinge) and add L_2 regularization on z' to prevent the magnitude explosion. Training uses FGSM for efficiency and separates gradients flowing through the energy and classification heads to the adversary, preventing energy awareness.

Despite its simplicity, our energy head detects adversarial examples produced by stronger PGD attacks and does not react to natural OOD data, provided that batch-independent normalization is used, so that training and evaluation compute energies consistently. However, we also observe the limitations: the classification robustness does not transfer and the energy values can explode when the hinge loss is used without regularization.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Contributions. (i) We propose a projection head yielding an energy score $E = \|z'\|^2$. (ii) We introduce the soft energy separation loss with L_2 regularization and analyze its stability. (iii) We implement FGSM training and FGSM+PGD-20 evaluation on CIFAR-9 with the 10th class as OOD, reporting AUROC and robust-after-rejection metrics. (iv) We demonstrate that batch-independent normalization is crucial for energy alignment between training and testing. (v) We demonstrate that our method does not mistake OOD for adversarial data. (vi) We report failure cases, such as non-transfer of classification robustness, and provide the details on head complexity and loss functions.

2 Related Work

Energy-based OOD detection. Liu et al. [2020] propose using the energy defined by the negative log partition function as a score for OOD detection and show that it reduces the false positive rate by 18% compared to the softmax confidence. Their framework allows energy to be used as a parameter-free inference score or as a trainable cost function with square hinge loss. We adapt such an idea, but use the squared norm of a projection instead of the logit-based energy and train the projection head jointly with classification.

Adversarial training and attacks. Adversarial training casts robustness as a saddle-point optimization problem and uses the inner maximization to generate worst-case perturbations. Madry et al. [2018] identify projected gradient descent (PGD) as a universal first-order adversary and demonstrate robust models on MNIST and CIFAR. The fast gradient sign method (FGSM) introduced by Goodfellow et al. [2015] provides an efficient way to generate adversarial examples by linearizing the loss around the input. Our training uses energy-blind FGSM, while evaluation includes FGSM and PGD-20. AutoAttack combines multiple attacks to reliably evaluate robustness and highlights that PGD may overestimate robustness; it recommends an ensemble of attacks as a minimal test.

Normalization layers and dataset shift. Batch normalization (BN, Ioffe and Szegedy [2015]) normalizes layer inputs using batch statistics to reduce internal covariate shift, improving training speed and acting as a regularizer. However, BN uses running estimates during evaluation, and mismatched statistics under distribution shift can harm performance. We find that using batch-independent normalization (e.g. instance normalization, Ulyanov et al. [2016]) is necessary to align energy distributions between train and test.

Uncertainty and distribution shift. Robustness under distribution shift and OOD inputs is necessary for safe deployment. Ovadia et al. [2019] benchmark predictive uncertainty methods and show that calibration in the i.i.d. setting does not translate to calibration under shift and that evaluating uncertainty under shift is more meaningful. Our method complements this line by focusing on detection via energy scores rather than calibration.

3 Method

3.1 Architecture and Energy Score

Let f_θ denote the backbone mapping an input image $x \in \mathbb{R}^d$ to a representation $z = f_\theta(x)$. We append a projection head f_η that maps z to the same or lower dimensional vector $z' = f_\eta(z) \in \mathbb{R}^k$. The classifier branch predicts the class probabilities from z via a linear layer and cross-entropy loss. The energy branch computes the score $E(z') = \|z'\|_2^2$ that we aim to make small for clean inputs and large for adversarial ones. In practice, f_η is a small multilayer perceptron.

3.2 Energy Separation Loss

Given a batch of clean examples $\{x_i\}$ and adversarial examples $\{x_i^{\mathrm{adv}}\}$ generated on the fly, we compute energies E_i and E_i^{adv} as described above. We minimize the total loss

$$\mathcal{L} = \frac{1}{B} \sum_i \Big[\underbrace{\mathrm{CE}(y_i, f_\theta(x_i))}_{\text{classification}} + \lambda_{\mathrm{sep}}\, \ell_{\mathrm{sep}}\big(E_i, E_i^{\mathrm{adv}}\big) + \lambda_2 \|z_i'\|_2^2 \Big], \tag{1}$$

where CE is the cross-entropy loss and ℓ_{sep} encourages separation between clean and adversarial energies. We experiment with three variants:

- **Hinge loss:** $\ell_{\text{sep}}(E, E^{\text{adv}}) = \max(0, \epsilon - E) + \max(0, E^{\text{adv}} - (\epsilon + \Delta))$. This penalty leads to energy explosion in practice unless λ_2 is tuned.
- **Softplus:** $\ell_{\text{sep}}(E, E^{\text{adv}}) = \text{softplus}(\epsilon - E) + \text{softplus}(E^{\text{adv}} - (\epsilon + \Delta))$, which is differentiable and alleviates gradient vanishing during training, making the joint training process more stable.
- **Squared hinge:** $\ell_{\text{sep}}(E, E^{\text{adv}}) = c \cdot \max(0, \epsilon - E)^2 + c \cdot \max(0, E^{\text{adv}} - (\epsilon + \Delta))^2$ that behaves similarly to softplus, provided that c is small enough to prevent initial penalty explosion.

We set ϵ as the maximum allowed value for clean energy and Δ as a margin hyperparameter. Regularization on z' prevents the projection from shrinking or exploding. During the adversarial example generation, we do *not* backpropagate through the energy branch, ensuring the attack is *energy-blind* and does not exploit our detector.

3.3 Adversarial Training and Evaluation

Adversarial training solves a saddle-point problem in which the inner maximization generates adversarial perturbations and the outer minimization updates the model parameters. We use FGSM for its efficiency and backpropagate only through the classification branch. The perturbations are constrained in the ℓ_∞ norm ball with radius $\varepsilon = 8/256$.

During evaluation, we generate adversarial examples using FGSM and 20-step PGD with step size $\alpha = 2/256$, both with doubled maximal allowed perturbation of $\varepsilon = 16/256$. Following Croce and Hein [2020], we sweep the threshold on the energy score to calculate area under the ROC curve (AUROC). We also report robust-after-rejection accuracy: classification accuracy over all the examples that did not exceed $\epsilon + \frac{\Delta}{2}$. OOD experiments treat the 10th class in CIFAR-10 as unknown and evaluate whether the energy rejects these inputs.

3.4 Normalization Alignment

During preliminary experiments, we observed that energy distributions for clean and adversarial examples behave differently between training and evaluation, often collapsing or even reversing. Investigation revealed that our backbone used batch normalization layers that adapt to batch statistics during training but use running estimates at evaluation. When adversarial examples dominated the batch, the running statistics drifted and corrupted the energy. To remedy this, we use instance normalization to perform exactly the same calculations both in train and test time. Figure 1illustrates how using IN stabilizes energy distributions.

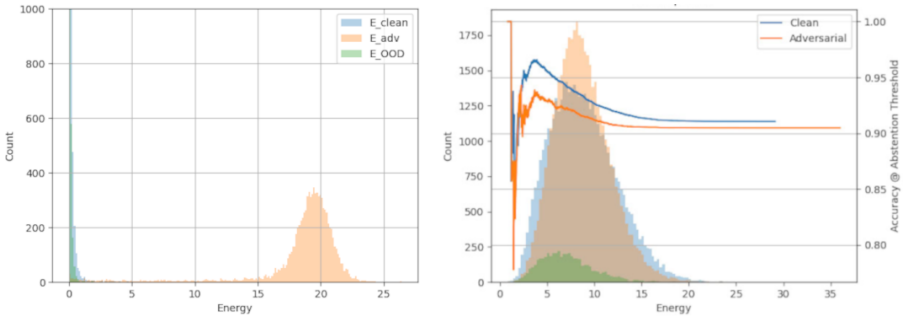

Figure 1: Energy histograms with instance (left) and batch (right) normalization. Under BN, clean and adversarial energies overlap. Using IN shifts adversarial energies higher and clean energies lower, enabling separation. Softplus and squared hinge losses achieve stable separation, whereas the hinge loss often causes uneven training and energy explosions unless regularization is used.

4 Experiments

We implement our method in PyTorch (https://anonymous.4open.science/r/manifold-projection-layer-B1DD/). The backbone is a pre-classification head ResNet-18 trained on CIFAR-9, i.e.

CIFAR-10 without class 10 ("truck"); the removed class serves as OOD data. The images are normalized and no data augmentation is used. We train for 5 epochs with batch size 16 using stochastic gradient descent with momentum 0.9 and learning rate 0.02. FGSM attacks use $\varepsilon = 8/255$. The hyperparameters λ_{sep} and λ_2 are tuned to facilitate smooth joint training without energy explosion or any part of loss dominating the training regime; we typically set $\lambda_{\text{sep}} = 1$ and $\lambda_2 = 5 \cdot 10^{-3}$. For evaluation, we generate 9,000 adversarial examples for each attack type and compute accuracy on clean and adversarial examples, AUROC in adversarial and OOD detection, and robust-after-rejection accuracy at the rejection threshold $\epsilon + \frac{\Delta}{2}$.

5 Results and Analysis

Detection versus classification. Table 1 summarizes the results. Energy-shaped projections achieve high AUROC for both FGSM and PGD attacks (> 0.99), even when classification accuracy on PGD examples is zeroed. This indicates that adversarial perturbations cause a predictable increase in the energy norm even if the classifier fails on the perturbed images. Energy separation therefore transfers to unseen attacks. However, the classification robustness does not transfer: a stronger adversary manages to nullify classification accuracy.

Table 1: Detection performance on CIFAR-9 test set. $\mathrm{AUROC}_{\mathrm{adv}}$ and $\mathrm{AUROC}_{\mathrm{OOD}}$ are measured for adversarial vs clean and OOD vs clean detection tasks respectively; $\mathrm{ACC}_{\mathrm{clean}}$ and $\mathrm{ACC}_{\mathrm{adv}}$ denote classification accuracy on clean and adversarial examples respectively; RAR is robust-after-rejection accuracy at $\epsilon + \frac{\Delta}{2}$ rejection threshold.

Adversary	$\mathrm{ACC}_{\mathrm{clean}}$	$\mathrm{ACC}_{\mathrm{adv}}$	$\mathrm{AUROC}_{\mathrm{adv}}$	$\mathrm{AUROC}_{\mathrm{OOD}}$	RAR
FGSM	0.6886	0.7821	1	0.5522	0.6886
PGD-20	0.6886	0	0.9976	0.5522	0.6597

OOD detection. When evaluating on the held-out CIFAR class, energy scores for OOD images closely match those of clean in-distribution examples. The AUROC for OOD versus clean detection is around 0.55, indicating near-random performance. Therefore, while energy-shaped projections do not replace standard OOD detection mechanisms, they might be compatible with these, since OOD data is not mistaken for adversarial. Figure 2 provides ROC curves as an illustration.

6 Limitations and Broader Impact

Our study has several limitations. First, we evaluate on CIFAR-like data; the results may not generalize to more complex domains or modalities. Second, training uses FGSM; while detection transfers to PGD, we have not evaluated energy-aware attacks or AutoAttack, which might circumvent our detector. Third, the projection head is tuned manually; automating its architecture and hyperparameters is left to future work. Finally, our method does not address distribution shift beyond adversarial perturbations: energy fails to detect unrelated OOD inputs. We encourage future work to evaluate compatibility of other OOD detection methods with energy-based projection heads.

7 Conclusion

We proposed an energy-shaped manifold projection head for adversarial detection. By training a projection head with a soft separation loss and regularizing the projected representation, we obtain a robust energy score that distinguishes adversarial inputs even when the classification robustness fails. Our experiments highlight the importance of normalization layer choice and show that softplus and squared hinge losses provide stable energy separation. At the same time, we report negative results: the method does not reject OOD data unrelated to the training distribution, and classification robustness does not improve. We hope our analysis and ablations will inspire further research into reliable detection mechanisms.

References

Alex Krizhevsky. Learning multiple layers of features from tiny images. pages 32–33, 2009. URL https://www.cs.toronto.edu/ kriz/learning-features-2009-TR.pdf.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 33, pages 18967–18979, 2020.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Sergey Ioffe and Christian Szegedy. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 448–456. JMLR.org, 2015.

Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *ArXiv*, abs/1607.08022, 2016. URL https://api.semanticscholar.org/CorpusID:16516553.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, pages 13991–14002, 2019.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning*, pages 2206–2216, 2020.

A Figures and Illustrations

A.1 Threshold sweep and rejection.

Figure 2 plots ROC curves for FGSM and PGD attacks. Energy-trained model is consistently good in adversarial detection, but is not suitable for OOD detection.

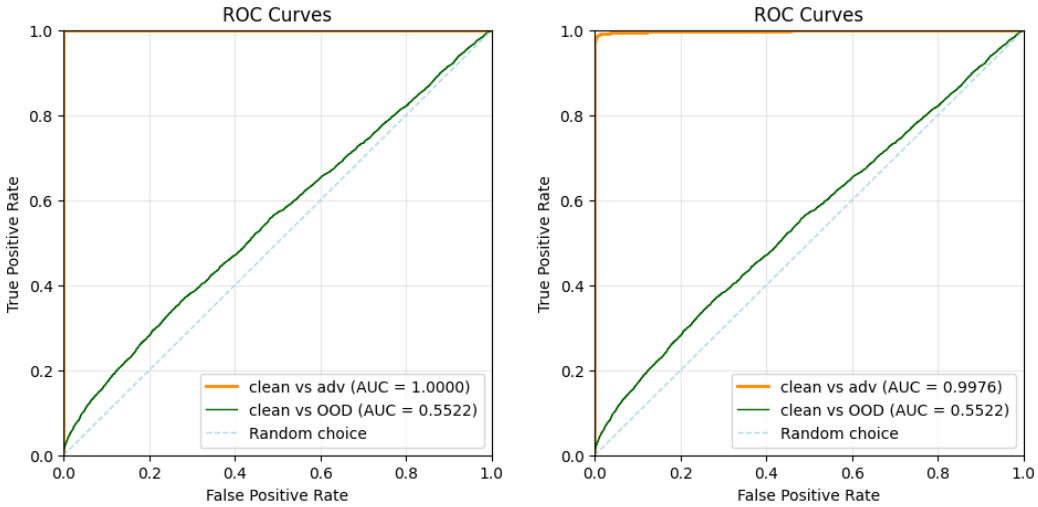

Figure 2: ROC curves for detecting adversarial examples using softplus energy loss. Performance under PGD-20 is slightly worse than under FGSM but remains high.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: the abstract and introduction are based either on the experiments or prior work, both described later.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: see the Limitations section.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [NA]

Justification: only the experimental results are reported.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: section Experiments provides a link to the repository with the code along with the full experiment setup description in the paper itself for reproducibility.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: section Experiments provides a link to the repository with the code along with the full experiment setup description in the paper itself for reproducibility.

Guidelines:

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

6. **Experimental setting/details**

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

Answer: [Yes]

Justification: also the parameters are provided in the repository.

Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [No]

Justification: these are replaced with the metrics and plots of choice in the main text and the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [No]

   Justification: the experiments can be reproduced in standard Google Colab computing environment.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: we conform to the Code and legal requirements to our best.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [NA]

    Justification: the work describes specifically the technical impact; we expect no negative societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [NA]

    Justification: we expect no such risk.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: we credit the original prior work to our best.

    Guidelines:

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

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: additionally, the code is anonymized to our best.

    Guidelines:

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

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

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

    Answer: [NA]

    Justification: the experiments include no human subjects.

    Guidelines:

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

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

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

    Answer: [NA]

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

    Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [NA]

Justification: only for the grammar purposes

Guidelines:

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