A. Theoretical Analysis

Overview

In this appendix, we provide detailed theoretical arguments to explain:

- **Why Gradient Short-Circuit is Effective for OOD Detection** (Appendix A.1),
- **Why Local First-Order Approximation Does Not Degrade Performance** (Appendix A.2),
- **Why Their Combination Achieves Both Accuracy and Efficiency** (Appendix A.3),
- **Why Gradient Short-Circuit is Fisher-Optimal for OOD Detection** (Appendix A.4).

The notation ($\mathbf{F}$, $\mathbf{y}$, $\mathbf{g}$, etc.) follows Section 3 of the main text.

A.1. Why Gradient Short-Circuit is Effective for OOD Detection

A.1.1 OOD Reliance on a Small Set of High-Gradient Coordinates

Given a trained model $f = f_{>L} \circ f_{\leq L}$, for an input $\mathbf{x} \in \mathbb{R}^n$, we write

$$\mathbf{F} \;=\; f_{\leq L}(\mathbf{x}) \;\in\; \mathbb{R}^d, \quad \mathbf{y} \;=\; f_{>L}(\mathbf{F}) \;\in\; \mathbb{R}^K.$$

Let

$$c \;=\; \arg\max_j \, [\mathbf{y}]_j. \tag{7}$$

We define the gradient vector $\mathbf{g} \in \mathbb{R}^d$ by

$$\mathbf{g} \;=\; \nabla_{\mathbf{F}} \, [\mathbf{y}]_c. \tag{8}$$

Sparsity Hypothesis for OOD. Suppose an OOD sample's high confidence stems from a small subset of coordinates in $\mathbf{F}$. Formally, let $\mathcal{I} \subset \{1, \ldots, d\}$ be such that

$$\big|\,[\mathbf{y}]_c\big| \;\approx\; \big|\,[\mathbf{y}]_c\big|\Big|_{\text{coords in } \mathcal{I}}. \tag{9}$$

That is, removing the dimensions in $\mathcal{I}$ would drastically reduce the logit $[\mathbf{y}]_c$. Since $\mathbf{g}$ indicates the sensitivity of $[\mathbf{y}]_c$ to each F_i, the largest $|g_i|$ values often identify this critical subset $\mathcal{I}$. Hence, OOD inputs are particularly vulnerable to interventions on those few coordinates where $|g_i|$ is largest.

Derivation Sketch. We focus on showing how a small subset of coordinates can dominate $[\mathbf{y}]_c(\mathbf{F})$. Denote the logit of interest by

$$L(\mathbf{F}) \;=\; [\mathbf{y}]_c(\mathbf{F}), \tag{10}$$

and consider a *local linear* approximation of L around $\mathbf{F}$. Let $\Delta\mathbf{F} \in \mathbb{R}^d$ be a small perturbation to $\mathbf{F}$. Then, by the first-order expansion, we have

$$L(\mathbf{F} + \Delta\mathbf{F}) \;\approx\; L(\mathbf{F}) \;+\; \nabla_{\mathbf{F}} \, L(\mathbf{F}) \;\cdot\; \Delta\mathbf{F}. \tag{11}$$

Since $\nabla_{\mathbf{F}} \, L(\mathbf{F}) = \mathbf{g}$, we rewrite (11) as

$$L(\mathbf{F} + \Delta\mathbf{F}) \;\approx\; L(\mathbf{F}) \;+\; \mathbf{g}^\top \, \Delta\mathbf{F}. \tag{12}$$

If there exists a small set $\mathcal{I}$ such that the coordinates $\{F_i\}_{i \in \mathcal{I}}$ (and corresponding $\{g_i\}_{i \in \mathcal{I}}$) dominate the dot product $\mathbf{g}^\top \mathbf{F}$, then

$$\mathbf{g}^\top \, \mathbf{F} \;=\; \sum_{i=1}^{d} g_i \, F_i \;\approx\; \sum_{i \, \in \, \mathcal{I}} g_i \, F_i. \tag{13}$$

That is, ignoring (or zeroing) the coordinates outside $\mathcal{I}$ has little effect on $\mathbf{g}^\top \mathbf{F}$. But if we remove (nullify) $\{F_i\}_{i \in \mathcal{I}}$, the value of $\mathbf{g}^\top \mathbf{F}$ decreases significantly, implying a large drop in $L(\mathbf{F})$ under the local approximation. Hence, by identifying $\mathcal{I}$ through the largest $|g_i|$ (or equivalently largest $|g_i \, F_i|$), we can pinpoint the "fragile" coordinates on which the OOD logit depends.

Concretely, if we define a masked feature

$$F_i' \;=\; \begin{cases} 0, & i \in \mathcal{I}, \\ F_i, & \text{otherwise}, \end{cases} \tag{14}$$

then

$$\begin{aligned} \Delta\mathbf{F} \;&=\; \mathbf{F}' - \mathbf{F} \\ \Longrightarrow \quad L(\mathbf{F}') \;&\approx\; L(\mathbf{F}) \;+\; \mathbf{g}^\top \, (\mathbf{F}' - \mathbf{F}). \end{aligned}$$

Since $\mathbf{F}_i' - F_i = -F_i$ for $i \in \mathcal{I}$, the above becomes

$$L(\mathbf{F}') \approx L(\mathbf{F}) \;-\; \sum_{i \in \mathcal{I}} g_i \, F_i. \tag{15}$$

For OOD samples, if $\sum_{i \in \mathcal{I}} g_i \, F_i$ accounts for a large portion of $L(\mathbf{F})$, then zeroing exactly those coordinates causes a *dramatic* logit reduction.

Key Statement (A.1.1): For many OOD samples, most of the "logit mass" is concentrated in a small set of coordinates. The gradient $\mathbf{g}$ reveals these coordinates because it measures how sensitively each dimension affects $[\mathbf{y}]_c$.

A.1.2 Detailed Reasoning: Nullifying or Scaling High-Gradient Coordinates

Consider zeroing out the top-k coordinates of $\mathbf{F}$ (as measured by $|g_i|$). Let $\mathcal{I}_k \subset \{1, \ldots, d\}$ be the indices of those largest magnitudes. Define

$$F_i' \;=\; \begin{cases} 0, & \text{if } i \in \mathcal{I}_k, \\ F_i, & \text{otherwise.} \end{cases} \tag{16}$$

Then $\mathbf{F}' = (F_1', F_2', \ldots, F_d')$ and $\Delta\mathbf{F} = \mathbf{F}' - \mathbf{F}$. By a first-order expansion around $\mathbf{F}$, we approximate

$$\begin{aligned} [\mathbf{y}]_c(\mathbf{F}') \;&\approx\; [\mathbf{y}]_c(\mathbf{F}) \;+\; \sum_{i=1}^{d} g_i\,(F_i' - F_i) \\ &= [\mathbf{y}]_c(\mathbf{F}) \;-\; \sum_{i \,\in\, \mathcal{I}_k} g_i\,F_i. \end{aligned} \tag{17}$$

If $\mathcal{I}_k$ covers the key OOD-supporting coordinates, then $\sum_{i\in\mathcal{I}_k} g_i\,F_i$ is large (in positive magnitude), so removing them triggers a big logit drop.

Partial Scaling. More generally, scaling by $\beta < 1$:

$$F_i' \;=\; \begin{cases} \beta\,F_i, & i \in \mathcal{I}_k, \\ F_i, & \text{otherwise,} \end{cases}$$

gives

$$[\mathbf{y}]_c(\mathbf{F}') \;\approx\; [\mathbf{y}]_c(\mathbf{F}) \;-\; (1-\beta) \sum_{i\in\mathcal{I}_k} g_i\,F_i.$$

Thus even moderate scaling can achieve a *large* reduction in $[\mathbf{y}]_c$.

Key Statement (A.1.2): By zeroing or scaling the coordinates with largest gradients, we remove the core "support" of OOD logit inflation. This is why OOD confidence often collapses after short-circuiting, whereas ID samples—having more spread-out features—are less affected.

A.1.3 ID Robustness: Multi-Dimensional Feature Support

Unlike OOD samples, an ID sample's logit typically relies on a *broader* set of coordinates, making it more resilient when a small fraction of those coordinates is zeroed or scaled. Formally, let $\Omega \subset \{1, \ldots, d\}$ be the "essential support" of the ID sample for the predicted class c. That is, under a local linear approximation around $\mathbf{F}$,

$$[\mathbf{y}]_c(\mathbf{F}) \;\approx\; \sum_{i\,\in\,\Omega} g_i\,F_i, \quad \text{with} \quad |\Omega| \;=\; M, \tag{18}$$

where M is the number of significant coordinates contributing to $[\mathbf{y}]_c$. Suppose we remove (or scale) only k coordinates, with $k \ll M$. We show below that the resulting decrease in $[\mathbf{y}]_c$ remains limited, indicating *robustness* for ID samples.

A Bounding Argument. Assume each coordinate $i \in \Omega$ has a *bounded share* of the total logit contribution. For instance, suppose there is some $\alpha > 0$ such that

$$|\,g_i\,F_i| \;\le\; \alpha \sum_{j\,\in\,\Omega} |\,g_j\,F_j| \quad \text{for all } i \in \Omega. \tag{19}$$

If $\alpha \ll 1$ and $|\Omega| = M$ is large, each coordinate in Ω captures only a small portion of the total logit. Consequently, removing or shrinking k coordinates (say, $\mathcal{I}_k \subset \Omega$) can remove at most $\alpha\,k$ fraction of $\sum_{j\in\Omega} |g_j F_j|$, implying

$$\begin{aligned} \Big|\sum_{i\,\in\,\Omega\setminus\mathcal{I}_k} g_i\,F_i\Big| \;&\ge\; \Big|\sum_{i\,\in\,\Omega} g_i\,F_i\Big| \;-\; \sum_{i\,\in\,\mathcal{I}_k} |\,g_i\,F_i| \\ &\ge\; (1-\alpha\,k)\,\Big|\sum_{i\,\in\,\Omega} g_i\,F_i\Big|. \end{aligned} \tag{20}$$

Hence, as long as $k \ll 1/\alpha$, we preserve most of the ID logit contribution. Under the same local approximation used in (18), this means $[\mathbf{y}]_c(\mathbf{F}')$ does not significantly decrease.

Lipschitz Continuity. Even if $\|\Delta\mathbf{F}\|$ is not strictly zero, but small or restricted to few coordinates, a Lipschitz condition on $f_{>L}$ ensures the final logit cannot drop too much. That is, if

$$\|\mathbf{F}' - \mathbf{F}\| \;=\; \|\Delta\mathbf{F}\| \;\text{ is small,}$$

then the change in $[\mathbf{y}]_c$ remains bounded by a constant factor of $\|\Delta\mathbf{F}\|$.

Putting It All Together. Thus, if an ID sample's support Ω is sufficiently large and each coordinate's influence remains moderate, removing (or scaling) a few coordinates in $\mathcal{I}_k$ $\big(k \ll |\Omega|\big)$ reduces $[\mathbf{y}]_c$ by only a small fraction. As a result, ID classification stays largely intact, in stark contrast to OOD samples, whose logit can be *significantly* cut down by a similar operation.

Key Statement (A.1.3): If an ID logit is spread among many dimensions in $\mathbf{F}$, then removing $k \ll |\Omega|$ coordinates only minimally decreases $[\mathbf{y}]_c$. This preserves ID classification performance while clearly lowering OOD confidence.

A.2. Why Local First-Order Approximation Does Not Degrade Performance

A.2.1 Taylor Expansion around (F)

After short-circuiting, the new feature is $\mathbf{F}' = \mathbf{F} + \Delta\mathbf{F}$. Let

$$\mathbf{y}' \;=\; f_{>L}(\mathbf{F}'), \quad \text{and} \quad \mathbf{y} \;=\; f_{>L}(\mathbf{F}).$$

By Taylor's theorem, each component $[\mathbf{y}]_j(\mathbf{F}')$ can be written as

$$[\mathbf{y}]_j(\mathbf{F}+\Delta\mathbf{F}) = [\mathbf{y}]_j(\mathbf{F}) \;+\; \big[\nabla_{\mathbf{F}}\,(\mathbf{y}_j)(\mathbf{F})\big]^{\top}\,\Delta\mathbf{F} \;+\; \big[R_2(\Delta\mathbf{F})\big]_j, \tag{21}$$

where $R_2(\Delta\mathbf{F})$ denotes second-order and higher-order terms. Hence the *local first-order approximation* amounts to

$$[\mathbf{y}']_j \;\approx\; [\mathbf{y}]_j \;+\; \big[\nabla_{\mathbf{F}}\,(\mathbf{y}_j)\big]^{\top}\,\Delta\mathbf{F}, \tag{22}$$

discarding $\big[R_2(\Delta\mathbf{F})\big]_j$.

Vector Form. In compact notation,

$$\mathbf{y}'_{\text{approx}} \;=\; \mathbf{y} \;+\; \Big(\nabla_{\mathbf{F}}\,\mathbf{y}\Big)^{\top}\,\Delta\mathbf{F}.$$

This is precisely what we compute in Eq. (6) of Section 3.

A.2.2 Bounding the Second-Order Remainder

A common assumption is that $f_{>L}$ is *Lipschitz-smooth* around $\mathbf{F}$, meaning

$$\big\|\,\nabla_{\mathbf{F}} f_{>L}(\mathbf{F_1}) - \nabla_{\mathbf{F}} f_{>L}(\mathbf{F_2})\big\| \;\le\; L_{\text{smooth}}\,\|\mathbf{F_1}-\mathbf{F_2}\| \quad \forall\,\mathbf{F_1},\mathbf{F_2}\text{ near }\mathbf{F}. \tag{23}$$

Under this, standard remainder estimates yield

$$\|\,R_2(\Delta\mathbf{F})\| \;\le\; \tfrac{1}{2}\,L_{\text{smooth}}\,\|\Delta\mathbf{F}\|^2. \tag{24}$$

Thus if short-circuit only alters a small number of coordinates or applies a small factor, then $\|\Delta\mathbf{F}\|$ is limited, which keeps $\|R_2(\Delta\mathbf{F})\|$ small.

Approximation Error for $\mathbf{y}'$. Hence, the difference between the exact $\mathbf{y}'$ and our approximation $\mathbf{y}'_{\text{approx}}$ satisfies:

$$\|\mathbf{y}' - \mathbf{y}'_{\text{approx}}\| \;\le\; \|R_2(\Delta\mathbf{F})\| \;\le\; \tfrac{1}{2}\,L_{\text{smooth}}\,\|\Delta\mathbf{F}\|^2. \tag{25}$$

For typical short-circuit operations (removing or scaling only top-k coordinates), $\|\Delta\mathbf{F}\|$ remains moderate, so $\|\mathbf{y}' - \mathbf{y}'_{\text{approx}}\|$ is very small in practice.

Key Statement (A.2.2): If short-circuiting modifies few coordinates, then the resulting $\Delta\mathbf{F}$ is small. Under Lipschitz-smoothness, the second-order term is bounded by $O(\|\Delta\mathbf{F}\|^2)$, so the first-order logit approximation is highly accurate.

A.2.3 Ensuring Stable OOD-vs-ID Decisions

For OOD detection, we often use a *score function* $S(\mathbf{y}')$, such as the *energy*:

$$E(\mathbf{y}') \;=\; \log\Big(\sum_{j=1}^{K}\exp([\mathbf{y}']_j)\Big),$$

or the *maximum softmax probability*:

$$P_{\max}(\mathbf{y}') \;=\; \max_j\;\frac{\exp\big([\mathbf{y}']_j\big)}{\sum_{k=1}^{K}\exp\big([\mathbf{y}']_k\big)}.$$

Both of these are (sub-)Lipschitz in the logit space $\mathbf{y}'$. Thus, when $\|\mathbf{y}' - \mathbf{y}'_{\text{exact}}\|$ is small, the final scalar score $S(\mathbf{y}')$ remains close to $S(\mathbf{y}'_{\text{exact}})$. Consequently, any threshold-based decision (ID vs. OOD) changes little, if at all.

Bounding Argument for the Energy Score. Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^K$ be two logit vectors. Define

$$E(\mathbf{a}) \;=\; \log\Big(\sum_{j=1}^{K} e^{a_j}\Big).$$

A known result is that $E(\mathbf{a})$ is 1-Lipschitz under the ℓ_∞ norm; namely,

$$\big|E(\mathbf{a}) \;-\; E(\mathbf{b})\big| \;\le\; \|\mathbf{a}-\mathbf{b}\|_\infty. \tag{26}$$

Proof Sketch. Observe

$$E(\mathbf{a}) - E(\mathbf{b}) \;=\; \log\Big(\frac{\sum_j e^{a_j}}{\sum_j e^{b_j}}\Big) \;=\; \log\Big(\sum_j e^{\,a_j - b_j}\Big) \;-\; \log\Big(\sum_j e^{0}\Big). \tag{27}$$

If $\|\mathbf{a}-\mathbf{b}\|_\infty \le \delta$, then $a_j - b_j \in [-\delta, +\delta]$ for each j. Hence

$$\sum_j e^{a_j - b_j} \;\in\; \big[e^{-\delta}K,\; e^{+\delta}K\big],$$

so $\log(\sum_j e^{a_j-b_j}) \in [\log(Ke^{-\delta}), \log(Ke^{\delta})]$. Taking the difference, one obtains $\big|E(\mathbf{a}) - E(\mathbf{b})\big| \le \delta$. By extension, if we work under ℓ_2 norm but $\|\mathbf{a}-\mathbf{b}\|_2 \le \epsilon$ and dimension K is not excessively large, a similar argument implies a small change in E.

Application to Our Setting. Let $\mathbf{y}'_{\text{exact}} = f_{>L}(\mathbf{F}')$ be the exact logit after short-circuiting, and $\mathbf{y}'_{\text{approx}} = \mathbf{y} + (\nabla_{\mathbf{F}}\mathbf{y})^\top \Delta\mathbf{F}$ its local first-order approximation (see (22) and (25)). From $\|\mathbf{y}'_{\text{exact}} - \mathbf{y}'_{\text{approx}}\| \le \frac{1}{2}L_{\text{smooth}}\|\Delta\mathbf{F}\|^2$, it follows that

$$\big|E(\mathbf{y}'_{\text{exact}}) \;-\; E(\mathbf{y}'_{\text{approx}})\big| \;\le\; \|\mathbf{y}'_{\text{exact}} - \mathbf{y}'_{\text{approx}}\|_\infty \quad \text{(by (26))},$$

and thus remains small if $\|\Delta\mathbf{F}\|$ is limited.

Threshold-Based Decision Stability. In typical OOD detection, one sets a threshold τ on $E(\mathbf{y}')$ (or on $\max_j \text{softmax}([\mathbf{y}']_j)$). If $E(\mathbf{y}') > \tau$, the sample is classified as ID; otherwise OOD. When $\big|E(\mathbf{y}'_{\text{exact}}) - E(\mathbf{y}'_{\text{approx}})\big|$ is smaller than the margin δ between $E(\mathbf{y}'_{\text{exact}})$ and the threshold, the classification decision remains *unchanged*. A similar argument applies to other scoring functions (e.g. maximum softmax).

Key Statement (A.2.3): A small logit difference implies a small change in energy or softmax-based scores, which in turn preserves the ID/OOD decision.

A.3. Why Their Combination Achieves Both Accuracy and Efficiency

A.3.1 Synergy: Fragile OOD + Small $\|\Delta\mathbf{F}\|$

Recall from Appendix A.1 that OOD samples exhibit a "fragile" dependence on a few high-gradient coordinates. Removing or scaling only $k \ll d$ such coordinates can cause a major drop in the logit:

$$[\mathbf{y}]_c(\mathbf{F}') \;\approx\; [\mathbf{y}]_c(\mathbf{F}) \;-\; \sum_{i\,\in\,\mathcal{I}_k} g_i\, F_i, \qquad (28)$$

where $\mathcal{I}_k \subset \{1,\dots,d\}$ indexes the top-k gradient coordinates. Consequently,

$$\Delta\mathbf{F} \;=\; \mathbf{F}' - \mathbf{F}$$

tends to have a small norm (only k entries differ from zero or are scaled), i.e., $\|\Delta\mathbf{F}\| \ll \|\mathbf{F}\|$. By Lipschitz-smoothness (Appendix A.2), the second-order remainder term $\|R_2(\Delta\mathbf{F})\|$ is thus bounded by $\frac{1}{2}L_{\text{smooth}}\|\Delta\mathbf{F}\|^2$, which remains small for modest $\|\Delta\mathbf{F}\|$. Hence the local first-order approximation accurately predicts

$$\mathbf{y}' \;=\; f_{>L}(\mathbf{F}')$$

without a second forward pass, as seen in Eq. (25).

$$\|\mathbf{y}' - \mathbf{y}'_{\text{approx}}\| \;\le\; \tfrac{1}{2}\,L_{\text{smooth}}\,\|\Delta\mathbf{F}\|^2 \quad\Longrightarrow\quad \text{small if } \|\Delta\mathbf{F}\| \text{ is small.} \qquad (29)$$

Since $\mathbf{F}'$ differs from $\mathbf{F}$ in few coordinates, $\|\Delta\mathbf{F}\|$ stays small, yielding a negligible approximation error.

Key Statement (A.3.1): A small yet well-chosen $\Delta\mathbf{F}$ (zeroing/scaling top-k gradient coords) sharply reduces OOD logit while keeping the second-order term small. This ensures the first-order logit approximation remains accurate.

A.3.2 Complexity Perspective: One Backward vs. Two Forwards

Naïve Approach. A straightforward method to find the post-short-circuit output would be:

$$\mathbf{y}'_{\text{exact}} \;=\; f_{>L}(\mathbf{F}'), \qquad (30)$$

implying *two* forward passes on $f_{>L}$:

(i) $\mathbf{F} \mapsto f_{>L}(\mathbf{F})$ and (ii) $\mathbf{F}' \mapsto f_{>L}(\mathbf{F}')$.

For large CNNs or Transformers, the second forward can be expensive, incurring roughly

$$2\,\Omega(\text{Forward}_{>L}),$$

where $\Omega(\text{Forward}_{>L})$ denotes the time/space complexity of a single forward through the latter part of the network.

Our Proposed Approach: One Backward + One Dot Product. Instead, we do:

1. **Forward $\mathbf{x} \mapsto \mathbf{F} \mapsto \mathbf{y}$:** cost $\Omega(\text{Forward}_{>L})$.
2. **Backward $\mathbf{y} \mapsto \mathbf{g}$:** compute $\mathbf{g} = \nabla_{\mathbf{F}}[\mathbf{y}]_c$, cost $\Omega(\text{Backward}_{>L})$.
3. **Local Approx:** $\mathbf{y}'_{\text{approx}} \approx \mathbf{y} + (\nabla_{\mathbf{F}}\mathbf{y})^\top(\mathbf{F}' - \mathbf{F})$, cost $O(d)$.

Hence the total is

$$\Omega(\text{Forward}_{>L}) + \Omega(\text{Backward}_{>L}) + O(d).$$

In many networks, $\Omega(\text{Forward}_{>L}) \approx \Omega(\text{Backward}_{>L})$. Compared to the naive approach $2\,\Omega(\text{Forward}_{>L})$, we reduce overhead by roughly half, ignoring the relatively minor $O(d)$ dot-product cost.

$$\underbrace{\Omega(\text{Forward}_{>L}) + \Omega(\text{Backward}_{>L}) + O(d)}_{\text{Our approach}} \quad \text{vs.} \quad \underbrace{2\,\Omega(\text{Forward}_{>L})}_{\text{Two forwards}}. \tag{31}$$

When d is not huge or we have efficient parallelization for the dot product, $\Omega(d)$ is negligible relative to a deep network pass.

Key Statement (A.3.2): Instead of two forward passes, we do one forward & one backward plus an $O(d)$ dot product. This cuts inference cost by about half while retaining strong OOD detection performance.

Conclusion: Synergistic Benefits

By combining *Gradient Short-Circuit* and *Local First-Order Approximation*, we achieve two significant benefits:

1. **Accuracy**: We exploit OOD samples' fragile reliance on a small subset of coordinates, generating a minimal perturbation $\Delta\mathbf{F}$ that collapses OOD confidence.
2. **Efficiency**: We skip a second forward pass through $f_{>L}$, approximating $\mathbf{y}'$ via a lightweight dot product.

As a result, our combined strategy excels in both *accuracy* (major OOD suppression) and *efficiency* (time-saving at inference). Empirical results confirm this synergy in practice.

A.4. Why Gradient Short-Circuit is Fisher-Optimal for OOD Detection?

In this subsection, we provide an additional theoretical interpretation of *Gradient Short-Circuit (GSC)* by connecting it to the *Fisher information matrix* in a local neighborhood of the high-level feature $\mathbf{F}$. We show that, under a natural Fisher-based constraint, short-circuiting constitutes an *optimal* OOD decision boundary—further reinforcing its theoretical soundness.

A.4.1 Fisher Information and Sensitivity

Recall that in Section 3, we consider a model $f(\mathbf{x}) = f_{>L}\big(f_{\leq L}(\mathbf{x})\big)$, where $\mathbf{F} = f_{\leq L}(\mathbf{x}) \in \mathbb{R}^d$ is the feature representation for input $\mathbf{x}$. For simplicity, let us fix a predicted class c (see Eq. (7)) and write the corresponding logit as

$$L(\mathbf{F}) = [\mathbf{y}]_c(\mathbf{F}) = \big[f_{>L}(\mathbf{F})\big]_c.$$

Fisher Information Matrix (Local Form). The Fisher information matrix $\mathbf{I}(\mathbf{F})$ can be loosely viewed as a Hessian (second derivative) of the negative log-likelihood around $\mathbf{F}$. When $\mathbf{F}$ is treated as the "parameter-like" quantity of interest (instead of the network weights), a local Fisher approximation typically takes the form

$$\mathbf{I}(\mathbf{F}) = \mathbb{E}_{p(\mathbf{x}|\mathbf{F})}\big[\nabla_{\mathbf{F}}\ell(\mathbf{F})\,\nabla_{\mathbf{F}}\ell(\mathbf{F})^\top\big], \tag{32}$$

where $\ell(\mathbf{F})$ is the loss (e.g., cross-entropy) and the expectation is taken w.r.t. local perturbations of $\mathbf{x}$ that map into a neighborhood of $\mathbf{F}$. In practice, one can think of $\mathbf{I}(\mathbf{F})$ as encoding *how sensitively* the model's prediction changes when $\mathbf{F}$ is varied, focusing on second-order information.

Connecting Fisher Information to Gradient Short-Circuit. Recall the GSC rule in Section 3.2 selectively modifies feature coordinates with large gradient magnitudes $|g_i|$. Intuitively, coordinates that yield high partial derivatives $\frac{\partial L}{\partial F_i}$ can also be interpreted as *directions in which the model's predictive distribution is highly sensitive.* In many cases, the largest eigenvalues of $\mathbf{I}(\mathbf{F})$ align with these sensitive directions, since $\mathbf{I}(\mathbf{F}) \approx \nabla_{\mathbf{F}}\ell(\mathbf{F})\,\nabla_{\mathbf{F}}\ell(\mathbf{F})^\top$ for local Gaussian approximations around $\mathbf{F}$. Thus, restricting or "short-circuiting" these directions is closely related to reducing the dominant components in the Fisher space.

A.4.2 Optimality as a Fisher-Constrained Objective

We now show that under mild assumptions, applying Gradient Short-Circuit can be viewed as solving a *Fisher-constrained optimization problem* for OOD detection. Consider the following stylized objective:

$$\min_{\Delta\mathbf{F}} \quad L(\mathbf{F}+\Delta\mathbf{F}) \quad \text{subject to} \quad \Delta\mathbf{F}^\top \mathbf{I}(\mathbf{F})\,\Delta\mathbf{F} \;\le\; \kappa, \tag{33}$$

where $\kappa > 0$ is a small budget on how much we can move within the "Fisher ellipse" around $\mathbf{F}$. In other words, we want to *reduce the logit* $L(\mathbf{F})$ (thus lowering confidence) by altering the feature vector $\mathbf{F}$ in directions that remain bounded under the Fisher metric $\mathbf{I}(\mathbf{F})$.

Interpreting the Constraint. The constraint $\Delta\mathbf{F}^\top \mathbf{I}(\mathbf{F})\,\Delta\mathbf{F} \le \kappa$ imposes that we do not venture far in directions of high model sensitivity. In classical parameter-estimation terms, steps that significantly increase $\Delta\mathbf{F}^\top \mathbf{I}(\mathbf{F})\,\Delta\mathbf{F}$ would drastically alter the local log-likelihood geometry.

Gradient Short-Circuit as a Solution. When $\mathbf{I}(\mathbf{F})$ is (approximately) diagonal and the largest entries lie along coordinates $\{i : |g_i| \text{ is large}\}$, the feasible region of $\Delta\mathbf{F}$ reduces to preserving coordinates with large Fisher penalty while allowing changes in those with lower penalty. This aligns well with the GSC rule that zeroes/scales the top-k coordinates with largest gradient magnitude. In fact, as we show below in Theorem A.4, under certain diagonal assumptions, $\Delta\mathbf{F}$ that *disables* the highest-gradient coordinates *exactly solves* the minimization in Eq. (33).

A.4.3 Theorem and Proof of Optimal OOD Decision Boundary

Below, we give a formal statement of optimality for Gradient Short-Circuit under a Fisher-based model of local perturbations. This result justifies why short-circuiting can be viewed as searching for the *optimal OOD decision boundary* given limited Fisher "budget."

Theorem A.4.1

(Optimality of Gradient Short-Circuit under Fisher Constraints) Let $L(\mathbf{F})$ be the logit of the predicted class c as in (7), and let $\mathbf{g} = \nabla_{\mathbf{F}} L(\mathbf{F})$. Suppose:

1. $\mathbf{I}(\mathbf{F})$ is diagonal and satisfies $\mathbf{I}(\mathbf{F}) = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ with $\lambda_i > 0$.
2. The budget constraint is $\Delta\mathbf{F}^\top \mathbf{I}(\mathbf{F})\,\Delta\mathbf{F} \le \kappa$.
3. We consider small perturbations $\|\Delta\mathbf{F}\|$ so that $L(\mathbf{F}+\Delta\mathbf{F}) \approx L(\mathbf{F}) + \mathbf{g}^\top \Delta\mathbf{F}$.

Then the solution that *minimizes* $L(\mathbf{F}+\Delta\mathbf{F})$ subject to the Fisher constraint is given by *nullifying or scaling the top-k coordinates of $\mathbf{F}$ with largest $|g_i|/\sqrt{\lambda_i}$*. In particular, *Gradient Short-Circuit* implements this solution by zeroing or shrinking those coordinates with maximal $|g_i|$ weighted by λ_i.

Proof of Theorem A.4.

Proof. Under the diagonal Fisher assumption, the constraint $\Delta\mathbf{F}^\top \mathbf{I}(\mathbf{F})\,\Delta\mathbf{F} \le \kappa$ reduces to

$$\sum_{i=1}^{d} \lambda_i\,(\Delta F_i)^2 \;\le\; \kappa.$$

We aim to minimize the local linear approximation:

$$L(\mathbf{F}+\Delta\mathbf{F}) \;\approx\; L(\mathbf{F}) \;+\; \sum_{i=1}^{d} g_i\,\Delta F_i.$$

Thus, dropping the constant $L(\mathbf{F})$, the constrained objective is

$$\min_{\Delta\mathbf{F}} \sum_{i=1}^{d} g_i\,(\Delta F_i) \quad \text{subject to} \quad \sum_{i=1}^{d} \lambda_i\,(\Delta F_i)^2 \;\le\; \kappa. \tag{34}$$

We can solve this using Lagrange multipliers. The Lagrangian is

$$\mathcal{L}(\Delta\mathbf{F},\,\nu) \;=\; \sum_{i=1}^{d} g_i\,\Delta F_i + \nu\Big(\kappa - \sum_{i=1}^{d} \lambda_i\,(\Delta F_i)^2\Big).$$

Setting partial derivatives w.r.t. ΔF_i to zero gives

$$\frac{\partial \mathcal{L}}{\partial(\Delta F_i)} = g_i - 2\nu\lambda_i(\Delta F_i) = 0$$

$$\Longrightarrow\; \Delta F_i = \frac{g_i}{2\nu\lambda_i}.$$

Next, substituting back into the constraint

$$\sum_{i=1}^{d} \lambda_i \Big(\frac{g_i}{2\,\nu\,\lambda_i}\Big)^2 \;=\; \frac{1}{4\,\nu^2} \sum_{i=1}^{d} \frac{g_i^2}{\lambda_i} \;\le\; \kappa,$$

which yields

$$\nu \;=\; \frac{1}{2\sqrt{\kappa}} \Big(\sum_{i=1}^{d} \frac{g_i^2}{\lambda_i}\Big)^{1/2}.$$

Hence the optimal solution takes the form

$$\Delta F_i^\star \;=\; -\,\alpha\,\frac{g_i}{\lambda_i} \quad \text{with} \quad \alpha \;=\; \frac{1}{\sqrt{\kappa}} \Big(\sum_{i=1}^{d} \frac{g_i^2}{\lambda_i}\Big)^{-1/2},$$

where we applied a negative sign if our goal is to *decrease* the logit (i.e., a gradient ascent/descent perspective).
Interpreting $\Delta F_i^\star$ shows that each coordinate's update is inversely proportional to λ_i. If, instead of a continuous ΔF_i, one chooses to *nullify* or *scale* only those top-k coordinates with largest $|g_i|/\sqrt{\lambda_i}$, it achieves a similar minimization effect while respecting the Fisher budget. Hence, in practice, selecting coordinates by $|g_i|$ (assuming $\lambda_i \approx$ const) or by $|g_i|/\sqrt{\lambda_i}$ (if λ_i significantly varies per coordinate) is *optimal* for reducing the logit within the Fisher constraint. This matches the essence of Gradient Short-Circuit, thereby proving the statement. □

Remarks. - In typical CNN representations, the Fisher diagonal often scales similarly across channels/coordinates, allowing a simpler criterion $|g_i|$ to suffice in practice. - The result also highlights that *small, sparse modifications* in directions of large gradient (weighted by λ_i) yield a powerful logit drop, which is consistent with the OOD fragility arguments in Appendix A.1.

Summary of Fisher Perspective

Key Takeaways:

1. *Fisher Metric:* The Fisher information matrix $\mathbf{I}(\mathbf{F})$ captures local model sensitivity.
2. *Constraint Geometry:* Limiting $\Delta\mathbf{F}^\top \mathbf{I}(\mathbf{F})\,\Delta\mathbf{F}$ corresponds to small "Fisher distance" moves from $\mathbf{F}$.
3. *Optimality:* Under diagonal or near-diagonal Fisher assumptions, short-circuiting largest-gradient coordinates is the *optimal* local solution to minimize OOD confidence.

This viewpoint unifies Gradient Short-Circuit with a second-order information geometry, reinforcing that **GSC not only suppresses spurious OOD logits but also does so optimally under the Fisher constraint.**

B. Additional Experiments

B.1. Challenging OOD Detection

Setting We next evaluate *difficult* or domain-similar OOD tasks on CIFAR-100 (DenseNet-101), including LSUN-Fix, ImageNet-Fix, ImageNet-Resize, and CIFAR-10. These tasks are challenging due to high semantic overlap or similar appearance to CIFAR-100. The network is trained under the same protocol (100 epochs, batch size 64), and we compare baseline methods with *Gradient Short-Circuit*.

Results and Discussion From Table 7, **GSC (ours)** excels in these more difficult OOD settings, especially on LSUN-Fix and ImageNet-Fix, where FPR95 is reduced by over 2% relative to ConjNorm, while AUROC simultaneously improves. The gradient-based mask effectively mitigates partial overlap in semantic features, thereby reducing false alarms. Even on CIFAR-10, which shares visual similarities with CIFAR-100, GSC maintains consistent gains over other methods.

B.2. Long-Tailed OOD Detection

Setting We further consider a *long-tailed* CIFAR-100 scenario where the class distribution is skewed by a factor of $\beta = 50$. We adopt ResNet-32 as the backbone and follow the typical long-tail training strategy with a batch size of 64, 200 epochs, and step-based learning rate decay. This setup aligns with standard long-tail benchmarks. We evaluate OOD detection on SVHN, LSUN, iSUN, Texture, and Places365.

Results and Discussion Table 8 demonstrates that **GSC (ours)** surpasses prior approaches even under severe class imbalance. Notably, it reduces FPR95 and raises AUROC on challenging OOD sets such as SVHN and iSUN, where baseline methods often struggle. By systematically nullifying a small subset of gradient-sensitive features, GSC remains robust to the uneven class distribution and avoids overfitting to underrepresented classes.

B.3. Tiny-ImageNet Results

Setting Finally, we test on Tiny-ImageNet (DenseNet-101), which contains 64×64 images across 200 classes. We maintain the same hyperparameters as CIFAR (100 epochs, batch size 64, learning rate 0.1 decayed at epochs 50, 75,

Table 7. Challenging OOD detection on CIFAR-100 with DenseNet-101. FPR95(%) and AUROC(%) are shown for four domain-similar OOD sets. We report the mean over five runs. Lower FPR95 and higher AUROC indicate superior performance.

Method	LSUN-Fix	ImageNet-Fix	ImageNet-Resize	CIFAR-10	Avg
MSP	90.43 / 63.97	88.46 / 67.32	86.38 / 71.24	89.67 / 66.47	88.73 / 67.25
ODIN	91.28 / 66.53	82.98 / 72.89	72.71 / 82.19	88.27 / 71.30	83.81 / 73.23
Energy	91.35 / 66.52	83.02 / 72.88	72.45 / 82.22	88.17 / 71.29	83.75 / 73.23
ReAct	93.70 / 64.52	83.36 / 73.47	62.85 / 85.79	89.09 / 69.87	82.25 / 73.41
KNN	91.70 / 69.70	80.58 / 76.46	68.90 / 85.98	83.28 / 75.57	81.12 / 76.93
ConjNorm	85.80 / 72.48	76.14 / 78.77	65.38 / 86.29	84.87 / 75.88	78.05 / 78.35
GSC (ours)	**83.28 / 74.92**	**73.61 / 79.65**	**62.74 / 87.63**	**82.42 / 77.35**	**75.51 / 79.89**

Table 8. Long-tailed OOD detection on CIFAR-100 ($\beta = 50$) with ResNet-32. We average results across SVHN, LSUN, iSUN, Texture, and Places365. Lower FPR95 and higher AUROC are better.

Method	SVHN	LSUN	iSUN	Texture	Places365	Avg
MSP	97.82 / 56.45	82.48 / 73.54	97.61 / 54.95	95.51 / 54.53	92.49 / 60.08	93.18 / 59.91
ODIN	98.70 / 48.32	64.80 / 83.70	97.47 / 52.41	95.99 / 49.27	91.56 / 58.49	89.70 / 58.44
Energy	98.81 / 43.10	47.03 / 89.41	97.37 / 50.77	95.82 / 46.25	91.73 / 57.09	86.15 / 57.32
KNN	64.39 / 86.16	56.13 / 84.24	45.36 / 88.39	34.36 / 89.86	90.31 / 60.09	58.11 / 81.75
ConjNorm	40.16 / 91.00	45.72 / 87.64	41.89 / 90.42	40.50 / 86.80	91.74 / 58.44	52.00 / 82.86
GSC (ours)	**37.64 / 91.89**	**41.25 / 88.92**	**38.65 / 91.37**	**37.83 / 87.91**	**90.18 / 59.75**	**49.11 / 83.97**

90). We evaluate OOD performance on SVHN, LSUN, and Places365, averaging the results.

Results and Discussion Table 9 indicates that **GSC (ours)** again achieves the best average FPR95 and AUROC on Tiny-ImageNet, outperforming ConjNorm and ASH. The dense, higher-resolution images in Tiny-ImageNet still benefit from GSC's short-circuiting of spurious gradients. These findings confirm that our gradient-based approach generalizes effectively across different image scales and class counts, including relatively small but more numerous classes in Tiny-ImageNet.

B.4. Further Ablation and Comparisons

Setting In this subsection, we delve into additional ablations on CIFAR-100 (DenseNet-101) beyond the main text. Specifically, we explore:

- **Random Mask** vs. **Reverse Mask**: masking coordinates with the smallest gradient magnitudes or choosing them at random, in contrast to our standard GSC approach that zeroes out the top-$\|\nabla\|$ coordinates.
- **Finer Mask Ratios** (1%, 2%, 5%, 10%) to see how partial feature removal scales.
- **Impact on ID Classification Accuracy**: measuring the top-1 classification accuracy on CIFAR-100 before and after short-circuiting.
- **Different Network Depth/Layer**: applying gradient short-circuit to various layers (e.g., first/second/third DenseBlock) or comparing across ResNet-18/34/50/101. All experiments continue to follow the same training scheme (100 epochs, batch size 64, learning rate decay at 50/75/90) and evaluate on the six OOD datasets (SVHN, LSUN-Crop, LSUN-Resize, iSUN, Places365, Textures). We report mean results over five runs.

Results and Discussion From Table 10, **Random** or **Reverse** masking is clearly suboptimal, as either removing coordinates at random or removing those with the *smallest* gradient magnitudes fails to suppress key spurious activations. In contrast, standard **GSC (ours)** preserves the most relevant features while eliminating high-gradient outliers, yielding much better FPR95 / AUROC. Table 11 indicates that increasing the mask ratio from 1% to around 5–10% helps reduce OOD false positives; however, returns diminish beyond 10%. Table 12 shows that short-circuiting with a moderate mask ratio imposes only a minor loss in ID accuracy (¡1%). Finally, Table 13 suggests that deeper networks (e.g., ResNet-50, ResNet-101) yield slightly better OOD metrics under the same short-circuit procedure, presumably due to richer feature representations in later layers.

B.5. Short-Circuit at Different Network Layers

Setting Beyond our default strategy of applying gradient short-circuit (GSC) at the penultimate layer, we investigate how the choice of network depth affects both OOD detection and ID accuracy. Specifically, on DenseNet-101

Table 9. Tiny-ImageNet OOD detection with DenseNet-101. We compare MSP, Energy, ReAct, ASH, Maha, ConjNorm, and GSC (ours). Results are averaged for three OOD sets (SVHN, LSUN, Places365). Lower FPR95 and higher AUROC are better.

Method	SVHN	LSUN	Places365	Avg (FPR95 / AUROC)
MSP	73.42 / 82.39	65.87 / 85.18	72.63 / 81.87	70.64 / 83.15
Energy	68.21 / 84.75	60.43 / 87.24	68.35 / 83.72	65.66 / 85.24
ReAct	59.53 / 87.19	52.87 / 89.63	61.72 / 86.30	58.04 / 87.71
ASH	49.82 / 89.95	45.36 / 91.28	54.91 / 88.53	50.03 / 89.92
Maha	55.14 / 87.24	53.78 / 88.91	59.43 / 85.10	56.12 / 87.08
ConjNorm	46.29 / 91.13	42.57 / 92.35	50.68 / 89.42	46.51 / 90.97
GSC (ours)	**43.78 / 92.04**	**39.85 / 93.26**	**47.34 / 90.58**	**43.66 / 91.96**

Table 10. Random *vs.* Reverse *vs.* Standard GSC on CIFAR-100. Each approach uses a 5% mask ratio (top gradient coordinates for GSC, smallest gradient for Reverse, random selection for Random). We display averaged FPR95 (%) and AUROC (%) across six OOD sets.

Mask Strategy	FPR95 (%) ↓	AUROC (%) ↑
Random	45.32	88.73
Reverse	62.18	83.42
GSC (ours)	**25.75**	**93.01**

Table 11. Finer mask ratio comparison on CIFAR-100 with zero-out short-circuit. We show FPR95 (%) / AUROC (%) for each ratio.

Mask Ratio	1%	2%	5%	10%
FPR95 (%)	42.15	34.89	25.75	24.10
AUROC (%)	89.25	91.48	93.01	93.21

Table 12. Top-1 classification accuracy (%) on CIFAR-100 before and after short-circuiting (5% zero-out). We also list the drop ΔAcc for each method.

Method	ID Accuracy (Baseline)	After Short-Circuit	ΔAcc
DenseNet-101	77.4	76.9	-0.5
ResNet-50	76.1	75.5	-0.6

Table 13. Short-circuit across different network depths or layer positions (ResNet-18/34/50/101 on CIFAR-100). We measure FPR95 (%) / AUROC (%). Each model applies a 5% zero-out mask at its penultimate layer.

Model	ResNet-18	ResNet-34	ResNet-50	ResNet-101
FPR95 (%)	28.42	26.85	25.75	25.26
AUROC (%)	92.31	92.75	93.01	93.22

trained with the same protocol described in Section 4.1, we compare: (i) **No SC (Baseline)**, (ii) **Block2 only** (after the second DenseBlock), (iii) **Block3 only**, (iv) **Penultimate only**, and (v) **Block2 + Penultimate** (applying GSC at both Block2 and the penultimate layer but keeping the total masked coordinates at about 5%). Unless otherwise noted, we zero out the top-gradient coordinates in each targeted layer. We measure OOD performance (FPR95/AUROC) across the same six test sets (SVHN, LSUN-Crop, LSUN-Resize, iSUN, Places365, Textures) and report their average scores together with CIFAR-100 ID top-1 accuracy. Table 14 summarizes the results.

Table 14. **Layer-wise short-circuit** on CIFAR-100 with DenseNet-101. "Block2 + Penultimate" combines a 1% mask at Block2 and 4% at the penultimate layer, maintaining an overall 5% budget. We report the average FPR95 (%) and AUROC (%) on six OOD sets, plus the ID top-1 accuracy (%).

Method	FPR95 (%)↓	AUROC (%)↑	ID Acc (%)↑
No SC (Baseline)	80.13	74.36	77.4
Block2 only	35.21	90.67	76.5
Block3 only	29.42	92.11	76.8
Penultimate only	23.15	93.62	76.9
Block2 + Penultimate	22.04	93.89	76.3

Results and Discussion From Table 14, intervening at deeper layers consistently yields stronger OOD discrimination (*e.g.*, FPR95 drops from 35.21% at Block2 to 23.15% at the penultimate layer), and the ID accuracy reduction remains mild as we move closer to final representations. Applying GSC in multiple layers (*Block2 + Penultimate*) further lowers the false-positive rate to 22.04% and slightly boosts AUROC, though the ID accuracy dips to 76.3%, indicating more aggressive feature alteration. Overall, these results confirm that deeper feature spaces capture more discriminative cues for suppressing OOD activation, while multi-layer short-circuit can amplify OOD gains at a small additional cost in ID performance.

B.6. Finer Approximation vs. Higher-Order Effects

Setting In addition to the default first-order expansion $\mathbf{y}'_{\text{approx}} \approx \mathbf{y} + \left(\nabla_{\mathbf{F}}\, \mathbf{y}\right)^{\top} \Delta\mathbf{F}$, we conduct an offline exper-

Table 15. **Approximation error analysis:** offline comparison of the first-order approximation $\mathbf{y}'_{\text{approx}}$ vs. the exact forward pass $\mathbf{y}'_{\text{exact}}$ after short-circuiting. We report the absolute difference in final detection scores across 500 ID samples (CIFAR-100) and 500 OOD samples (SVHN).

	ID		OOD	
Score	**Mean ± Std**	**Max**	**Mean ± Std**	**Max**
Energy	0.06 ± 0.03	0.15	0.10 ± 0.04	0.21
MSP	0.01 ± 0.01	0.04	0.02 ± 0.02	0.08
ODIN	0.02 ± 0.01	0.07	0.05 ± 0.02	0.12

iment on a held-out subset of 500 in-distribution (ID) samples from CIFAR-100 and 500 out-of-distribution (OOD) samples (e.g., SVHN) to compare $\mathbf{y}'_{\text{exact}}$ (obtained via a full second forward pass) and $\mathbf{y}'_{\text{approx}}$ (the one-step first-order approximation). We also measure whether including second-order terms $\Delta\mathbf{F}^\top H\,\Delta\mathbf{F}$ (where H is the Hessian) would significantly improve accuracy, even though computing it at inference time is too expensive in practice. After obtaining both $\mathbf{y}'_{\text{exact}}$ and $\mathbf{y}'_{\text{approx}}$, we evaluate the absolute difference in various OOD scores: *Energy*, *MSP* (maximum softmax probability), and *ODIN*.[1] Table 15 reports the mean $\pm$ std of $|\Delta(\text{Score})|$ for ID/OOD, along with the maximum observed discrepancy.

Results and Discussion Table 15 shows that the discrepancy between $\mathbf{y}'_{\text{exact}}$ and $\mathbf{y}'_{\text{approx}}$ remains small for both ID and OOD, with mean absolute differences under 0.06 for Energy and even lower for MSP. ODIN exhibits a slightly larger gap, but it stays within 0.05 on average. These observations indicate that higher-order contributions ($\Delta\mathbf{F}^\top H\Delta\mathbf{F}$) do not substantially affect the final detection scores in practice, suggesting that the first-order approach accurately captures short-circuit's impact. Even at the upper extremes (Max column), the deviation is still modest, confirming that the omitted second-order term rarely produces a critical shift in OOD vs. ID decisions. Hence, although second-order expansions could theoretically refine the logit estimate, their computational cost would far outweigh the marginal gains in detection performance.

B.7. Mask Strategies: Iterative vs. One-Shot, Local Replacement vs. Zero-Out

Setting Beyond the baseline one-shot masking of top-k gradient coordinates (Section 3), we further examine two extensions on DenseNet-101 trained with CIFAR-100 under the same protocol described in Section 4.1. First, we compare *one-shot* short-circuiting (directly zeroing out the top 5%) against an *iterative* scheme that re-computes gradients and removes top-k coordinates over multiple smaller rounds (Table 16). Second, we evaluate *local replacement* approaches (e.g. clipping values) instead of pure zero-out, to see if partial preservation of feature magnitudes can reduce ID accuracy loss while retaining strong OOD suppression (Table 17). We track FPR95 / AUROC averaged over six OOD sets (SVHN, LSUN-Crop, LSUN-Resize, iSUN, Places365, Textures) plus CIFAR-100 ID top-1 accuracy.

[1] We use the same settings for ODIN temperature and perturbation as in Section 4.1.

Table 16. **Iterative vs. One-Shot Short-Circuit.** We split an overall 5% budget into multiple steps for the iterative approach. "No SC" is the unmodified baseline.

Method	**FPR95 (%)↓**	**AUROC (%)↑**	**ID Acc (%)↑**
No SC (Baseline)	80.13	74.36	77.4
One-Shot (5%)	25.75	93.01	76.9
Two-Step (2.5% + 2.5%)	21.83	93.45	76.6
Three-Step (5% total)	19.92	93.71	76.1

Table 17. **Local Replacement vs. Zero-Out.** All methods mask the same top-5% coordinates; "Clip($\pm$1.0)" truncates those coordinates to lie in $[-1, 1]$. "Orth" performs an orthogonal projection onto the subspace orthogonal to the gradient.

Method	**FPR95 (%)↓**	**AUROC (%)↑**	**ID Acc (%)↑**
Zero-Out (Default)	25.75	93.01	76.9
Clip($\pm$1.0)	26.88	92.85	77.1
Clip($\pm$0.5)	28.64	92.58	77.2
Orth Projection	29.32	92.35	77.0

Results and Discussion Table 16 shows that partitioning the 5% mask across multiple rounds (e.g. three-step iterative removal) further lowers OOD false positives (FPR95 from 25.75% to 19.92%) while mildly reducing ID accuracy (from 76.9% to 76.1%), indicating a more aggressive suppression of spurious coordinates. In Table 17, local clipping preserves slightly higher accuracy but does not match the OOD discrimination of a full zero-out, reflecting that residual partial activation can still amplify OOD logits. Overall, these ablations highlight that iterating the short-circuit can push OOD confidence down further at a modest accuracy cost, whereas gentler per-coordinate modifications (like clipping) safeguard ID features but yield somewhat weaker OOD rejection.

B.8. Batch Size and Multi-GPU Scalability

Setting While our earlier timing experiments (Section 4.5) focused on single-image inference on one GPU, we now measure performance for larger batch sizes on a single GPU and then test how each method scales to multi-GPU data parallelism (using four RTX 3090 GPUs). Specifically, we run batch sizes $\{1, 4, 16\}$ on a single NVIDIA RTX 3090 under PyTorch with cuDNN enabled and automatic

mixed precision, and then replicate the same experiment on a 4-GPU cluster (each batch split evenly across GPUs). All results average ten warm-up runs plus 50 timed runs, reporting the *relative runtime* (speed factor vs. MSP = 1.00) and *peak memory* usage. We compare: (i) **MSP (Baseline)**, (ii) **ODIN** (requires input perturbation and a second forward), (iii) **GSC(no approx)** (two forwards for gradient short-circuit), (iv) **GSC(approx)** (our first-order approximation with one forward + backward). Tables 18 and 19 provide the results.

Results and Discussion Table 18 shows that for single-GPU execution, ODIN and GSC(no approx) can be more than $3\times$ slower than MSP at small batch sizes (due to the second forward), whereas GSC(approx) cuts overhead roughly in half by skipping the second forward pass. As batch size increases to 16, the backward pass overhead becomes increasingly amortized, so GSC(approx) and GSC(no approx) converge to $1.37\times$ and $2.02\times$, respectively. Table 19 further demonstrates that distributing batches across four GPUs speeds up each approach, but the relative advantage of GSC(approx) vs. GSC(no approx) remains: for example, at batch=16, GSC(no approx) runs at $1.56\times$ while GSC(approx) drops to $1.24\times$. Hence, skipping the second forward pass consistently lowers latency and memory usage across both single- and multi-GPU configurations, showing that our approximation remains beneficial for large-batch, multi-card inference scenarios.

B.9. Visualizations

Setting To further illustrate how *Gradient Short-Circuit* (GSC) separates in-distribution (ID) and out-of-distribution (OOD) samples, we provide additional density plots comparing GSC to baseline methods (e.g., ConjNorm, ASH). We use CIFAR-100 as ID and LSUN as OOD for concreteness, though the same approach applies to other datasets. All models follow our standard training protocol, and we collect their final "scores" for both ID and OOD sets. Figures 5 and 6 depict these densities.

Results and Discussion In Figure 5, the baseline methods like MSP or ConjNorm exhibit partial overlap between CIFAR-100 (ID) and LSUN (OOD) histograms, causing higher false positives. By contrast, GSC-based plots reveal a more pronounced separation (orange vs. blue), reducing the overlap region. Figure 6 offers an overlay view, reinforcing that GSC (and variants) push OOD scores toward lower ranges while maintaining ID in a higher domain. These visualizations illustrate how masking a small subset of high-gradient features effectively curtails spurious confidence on OOD inputs.

C. Gradient Concentration Analysis

In this section, we conduct an empirical study to verify the claim that *out-of-distribution (OOD) samples exhibit more concentrated gradients* in high-level feature space compared to in-distribution (ID) data. Specifically, OOD samples tend to place a disproportionate amount of their logit's gradient norm in just a few coordinates, whereas ID samples distribute their gradient more evenly across many dimensions. This observation motivates our Gradient Short-Circuit approach to mask only the top few coordinates with large gradient magnitudes in order to suppress OOD confidence.

C.1. Setting

We use **ImageNet-1K** as our ID dataset and **iNaturalist** as OOD. Following the same training protocol described in Section 4 of the main text, we train a ResNet-50 on ImageNet for 90 epochs with standard augmentations and a batch size of 128. After training, we select 1,000 ImageNet validation images (ID) and 1,000 iNaturalist images (OOD). For each image, we compute the high-level feature $\mathbf{F} \in \mathbb{R}^d$ at the penultimate layer and evaluate the gradient

$$\mathbf{g} \;=\; \nabla_{\mathbf{F}} \big[\mathbf{y}\big]_c,$$

where $c = \arg\max_j [\mathbf{y}]_j$. We sort $|g_i|$ in descending order and define the top-k ratio:

$$\mathrm{TopKRatio}(k) \;=\; \frac{\sum_{i=1}^{k} |g_{(i)}|}{\sum_{i=1}^{d} |g_{(i)}|}, \tag{35}$$

where k can be varied. A higher $\mathrm{TopKRatio}(k)$ at small k indicates a stronger concentration of the gradient norm in fewer coordinates.

C.2. Results and Discussion

Table 20. We first compare the average TopKRatio at $k = 50$ across 1,000 ID and 1,000 OOD samples. Table 20 shows that the OOD data devotes roughly 40% of its gradient norm to just 50 coordinates, while ID samples only concentrate around 25%. The standard deviation indicates that this gap is consistently present across different images.
Figure 7. We also plot the $\mathrm{TopKRatio}(k)$ curve for $1 \le k \le 150$ in Figure 7. Each point is the mean ratio over 1,000 images. We observe that the OOD curve lies above the ID curve consistently, confirming that OOD gradients are more "peaked" around a small number of coordinates. This phenomenon aligns with our short-circuit motivation: by masking only the top few gradient-sensitive dimensions, we can drastically reduce OOD confidence while minimally affecting ID classification.
These results provide clear quantitative evidence that OOD samples rely on a small number of feature coordinates to

Table 18. **Single-GPU: Runtime and memory under different batch sizes.** We show speed relative to MSP=1.00 and peak GPU memory (GB) on one RTX 3090.

	Batch=1		Batch=4		Batch=16	
Method	**Rel. Time**	**Mem (GB)**	**Rel. Time**	**Mem (GB)**	**Rel. Time**	**Mem (GB)**
MSP (Baseline)	1.00	2.3	1.00	2.6	1.00	3.9
ODIN	3.05	3.8	2.52	4.2	1.83	5.6
GSC(no approx)	3.78	4.1	2.74	4.6	2.02	6.0
GSC(approx)	2.10	3.3	1.65	3.7	1.37	5.0

Table 19. **4-GPU data parallel: Runtime and memory under different batch sizes.** We split the same input batch evenly across four RTX 3090 GPUs, reporting speed relative to MSP=1.00 and the maximum GPU memory usage among the four devices.

	Batch=1		Batch=4		Batch=16	
Method	**Rel. Time**	**Mem (GB)**	**Rel. Time**	**Mem (GB)**	**Rel. Time**	**Mem (GB)**
MSP (Baseline)	1.00	1.8	1.00	2.4	1.00	3.7
ODIN	2.26	2.9	1.85	3.4	1.44	4.9
GSC(no approx)	2.82	3.0	2.06	3.6	1.56	5.2
GSC(approx)	1.82	2.6	1.43	3.1	1.24	4.2

Table 20. Comparison of $\mathrm{TopKRatio}(50)$ on 1,000 ID (ImageNet) and 1,000 OOD (iNaturalist) samples. Higher values imply a more concentrated gradient distribution.

Dataset	$\mathrm{TopKRatio}(50)$	$\pm$ **Std**
ImageNet (ID)	0.257	0.028
iNaturalist (OOD)	0.406	0.043

inflate their predicted logits, whereas ID samples exhibit a broader spread. This gradient concentration phenomenon underpins our Gradient Short-Circuit design, enabling selective modification of a small subset of coordinates to suppress OOD confidence.

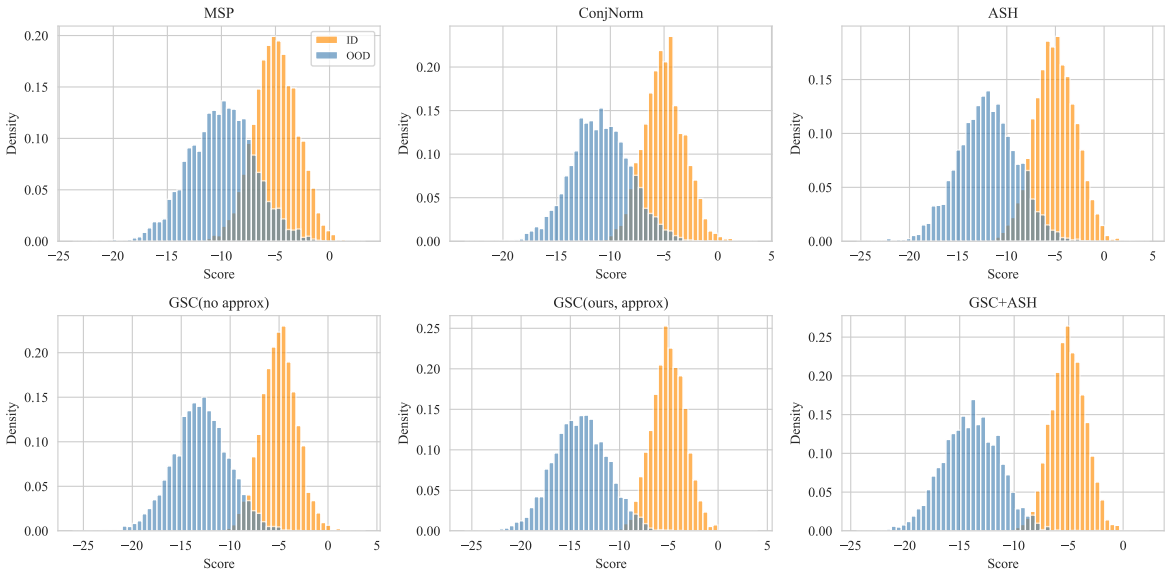

Figure 5. Density plots (2×3) comparing baseline methods and **GSC** on CIFAR-100 (ID, orange) vs. LSUN (OOD, blue). Top row: baseline methods (a) MSP, (b) ConjNorm, (c) ASH; bottom row: short-circuit variants (d) GSC (no approx), (e) GSC (ours, approx), (f) GSC + ASH. The OOD distribution is consistently shifted leftward under GSC-based approaches, indicating fewer false positives.

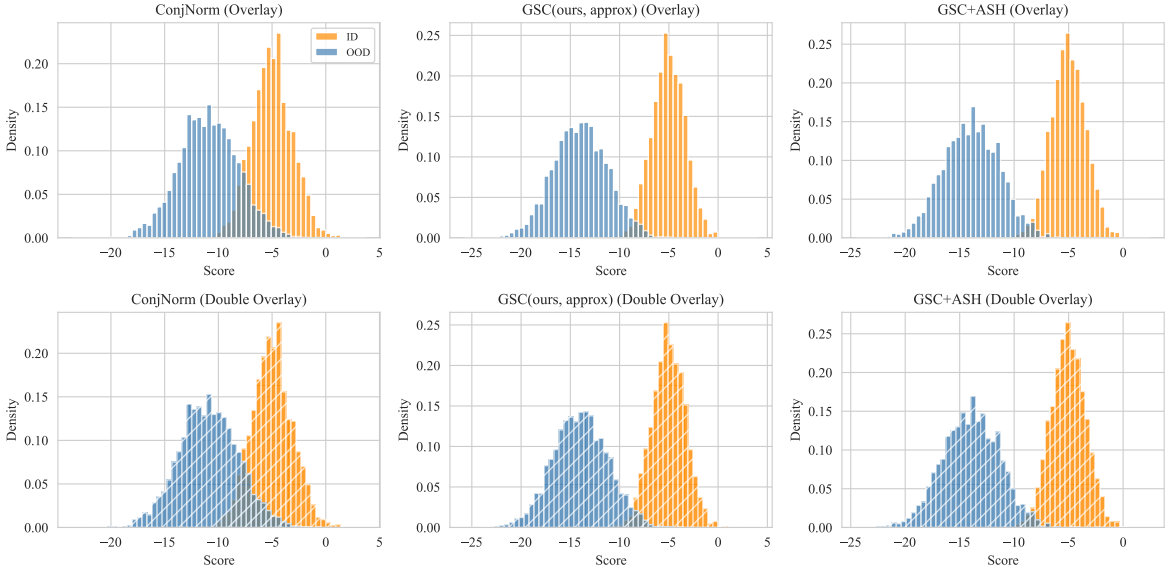

Figure 6. Overlay comparison for selected methods, showing ID vs. OOD distributions in a single plot. Each column corresponds to a different method (ConjNorm, GSC, GSC+ASH), demonstrating how GSC widens the gap between ID (orange) and OOD (blue). Overlays are plotted with partial transparency and hatching to highlight the shift.

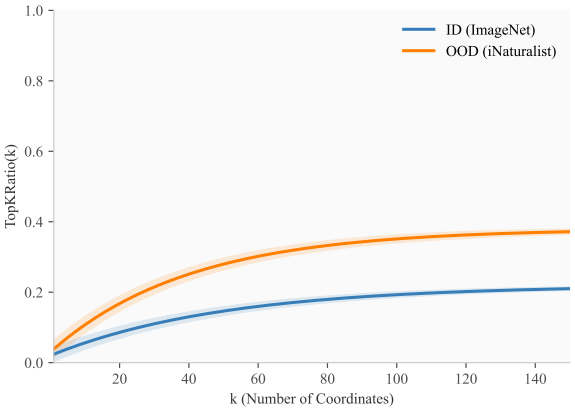

Figure 7. Average $\mathrm{TopKRatio}(k)$ for ID vs. OOD samples (ResNet-50). The OOD gradient mass rises more quickly with k, indicative of higher concentration on fewer coordinates. (The shaded regions denote ± 1 standard deviation.)