START SMART: LEVERAGING GRADIENTS FOR ENHANCING MASK-BASED XAI METHODS

Buelent Uendes, Shujian Yu & Mark Hoogendoorn
Vrije Universiteit Amsterdam
The Netherlands
{b.uendes}@vu.nl

ABSTRACT

Mask-based explanation methods offer a powerful framework for interpreting deep learning model predictions across diverse data modalities, such as images and time series, in which the central idea is to identify an instance-dependent mask that minimizes the performance drop from the resulting masked input. Different objectives for learning such masks have been proposed, all of which, in our view, can be unified under an information-theoretic framework that balances performance degradation of the masked input with the complexity of the resulting masked representation. Typically, these methods initialize the masks either uniformly or as all-ones. In this paper, we argue that an effective mask initialization strategy is as important as the development of novel learning objectives, particularly in light of the significant computational costs associated with existing mask-based explanation methods. To this end, we introduce a new gradient-based initialization technique called StartGrad, which is the first initialization method specifically designed for mask-based post-hoc explainability methods. Compared to commonly used strategies, StartGrad is provably superior at initialization in striking the aforementioned trade-off. Despite its simplicity, our experiments demonstrate that StartGrad enhances the optimization process of various state-of-the-art mask-explanation methods by reaching target metrics faster and, in some cases, boosting their overall performance.

1 INTRODUCTION

As machine learning models become deeply integrated into critical areas such as healthcare or medicine, the need for transparency and interpretability grows ever more urgent (Vellido, 2020). Explainable AI (XAI) research has responded by developing various XAI frameworks, with saliency methods—also referred to as feature attribution methods—at the forefront. These methods seek to highlight the most relevant inputs that drive a model's prediction, offering insights into how complex models, often regarded as black boxes, make decisions.

Mask-based explanation methods, a particularly powerful subset of feature attribution techniques, aim to learn sparse masks that highlight the key inputs driving a prediction. These methods are highly flexible: the objective function can be designed to capture specific notions of relevance, while constraints like sparsity or smoothness can be added to meet criteria for a 'good' explanation. This is often formalized through an information-theoretic framework, such as the information bottleneck (IB) principle (Tishby et al., 1999) and the rate-distortion function (Thomas & Joy, 2006). The carefully designed objective function in mask-based methods allows them to adapt effectively to different data modalities, resulting in superior performance compared to approaches such as Saliency maps (Simonyan et al., 2014), Integrated Gradients (Sundararajan et al., 2017), SmoothGrad (Smilkov et al., 2017), or surrogate-based methods like LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017). These approaches have been also shown to produce explanations that are fragile and noisy, or even ones that can be manipulated (Adebayo et al., 2018; Ghorbani et al., 2019; Slack et al., 2020).

Despite their effectiveness, state-of-the-art mask-based explanation methods come with heavy computational costs due to their prolonged optimization process with hundreds of epochs of iterations, especially when compared to gradient-based saliency techniques like SmoothGrad (Smilkov et al., 2017) or Integrated Gradients (Sundararajan et al., 2017). For instance, the most advanced mask-based methods in the vision domain such as the recently proposed WaveletX (Kolek et al., 2023) and ShearletX (Kolek et al., 2023), require orders of magnitude more execution time to generate explanations than their gradient-based counterparts.

This trade-off between performance and speed presents a major limitation for mask-based methods, especially in time-sensitive applications where rapid decisions are critical. As a result, users are often forced to choose between the superior faithfulness of mask-based explanations and the faster, but potentially less reliable, gradient-based alternatives. This tension between explanation accuracy and real-time applicability remains a key challenge in making mask-based methods more widely applicable in high-stakes environments such as healthcare.

To address the challenge of balancing performance and computational efficiency in mask-based methods, we draw on two key insights: First, while the way how to initialize masks in existing mask-based XAI methods is usually neglected, it plays a crucial role in optimization in terms of both running time and the final achievable maximum or minimum value. Second, although gradient-based saliency methods are not explicitly designed to meet desiderata of high-quality explanations, they do provide valuable signals about the model's decision-making process with minimal computational overhead. By combining these two insights, we propose StartGrad, a novel gradient-based mask initialization technique specifically designed for post-hoc explanation methods. StartGrad leverages gradient signals to provide provably superior initialization masks in terms of minimal distortion and sparsity—two essential criteria for mask-based explanation methods—compared to commonly used strategies. By doing so, StartGrad harnesses the strengths of gradient-based approaches to enhance existing mask-based explainability techniques.

We summarize our contributions as below:

- We introduce StartGrad, a novel gradient-based mask initialization algorithm grounded in the rate distortion explanation (RDE) framework (Macdonald et al., 2019), representing the first initialization technique explicitly designed to enhance the performance of mask-based explanation methods.
- We prove that StartGrad is superior at initialization compared to other initialization strategies in reducing distortion and improving sparsity—two essential criteria for effective mask-based explanations.
- Extensive experiments on vision and time-series tasks demonstrate that StartGrad enables state-of-the-art methods like ShearletX (Kolek et al., 2023) and ExtremalMask (Enguehard, 2023) to reach target metrics faster while also improving overall performance in some cases.
- To the best of our knowledge, this work presents the first comprehensive theoretical and empirical analysis of mask initialization techniques across both vision and time-series domains, providing critical insights for improving mask-based explanation methods.

2 Related Work

Feature attribution methods can be divided into white-box, gray-box, and black-box approaches, depending on the amount of information required to generate an explanation (Muzellec et al., 2024). Mask-based explanation methods are considered black-box attribution techniques, as they do not require access to the model's internal architecture and rely only on model predictions. Recently, these methods gained attraction in the community and have been subsequently applied to different domains such as vision (Fong & Vedaldi, 2017; Petsiuk et al., 2018; Kolek et al., 2022; 2023) or time series (Crabbé & Van Der Schaar, 2021; Enguehard, 2023; Liu et al., 2024; Zichuan Liu, 2024). A key advantage of these techniques is their explicit optimization of an objective function that formalizes desiderata for "good" explanations, such as parsimony and fidelity to the model and often formalized via the IB principle (Tishby et al., 1999) or the RDE framework (Macdonald et al., 2019) which formalize the trade-off between explanation complexity and predictive accuracy.

Despite their strong performance, mask-based approaches are often computationally expensive due to iterative optimization, limiting their practicality in real-world applications. However, much of the previous work has mainly investigated designing novel objective functions suited for either the data modality at hand (Ying et al., 2019; Crabbé & Van Der Schaar, 2021), proposing new architectural design choices such as learning the perturbation function(Enguehard, 2023; Liu et al., 2024), or even learning a trainable masking model that can produce masks after training (Dabkowski & Gal, 2017). To the best of our knowledge, no previous work has investigated the role of mask initialization scheme itself, which is surprising given the role initialization schemes can have on the overall performance of the optimization result (Glorot & Bengio, 2010; He et al., 2015). In fact, we noticed that across data modalities, different initialization techniques are used. For instance, mask-based XAI methods in the time-series domain initialize their masks uniformly at start (Crabbé & Van Der Schaar, 2021; Enguehard, 2023; Liu et al., 2024)[1], whereas vision explanation models use an all-ones initialization scheme (Kolek et al., 2022; 2023).

3 Background

Different objectives have been developed to learn an "optimal" mask $\mathbf{m}$. Based on our observation, these objectives can be unified under the same umbrella, formalizing the trade-off between the complexity of the masked input and its predictive performance from an information-theoretic perspective. At their core, all methods revolve around these two fundamental concepts, with additional constraints, such as smoothness, often incorporated to further refine the explanations.

Given a pre-trained predictive model $\Phi_c : \mathbb{R}^d \to [0,1]^c$ for a total of c classes and an input instance $\mathbf{x} \in \mathbb{R}^d$, the goal of the mask-based post-hoc explanation to a black-box output $\Phi_c(\mathbf{x})$ is to identify a sparse mask $\mathbf{m} \in \{0,1\}^d$ over $\mathbf{x}$, such that the resulting perturbed input $\tilde{\mathbf{x}}$ leads to the minimum performance drop in the model's prediction (Fong & Vedaldi, 2017; Macdonald et al., 2019). Usually, $\tilde{\mathbf{x}} := \mathbf{m} \odot \mathbf{x} + (1-\mathbf{m}) \odot \mathbf{u}$, where $\mathbf{u} \in \mathbb{R}^d$ is random noise following a predefined probability distribution V, such as a Gaussian distribution.

3.1 The Rate-Distortion Explanation (RDE) Framework

Formally, the objective of the RDE framework can be defined as the following constrained optimization problem (Macdonald et al., 2019; Kolek et al., 2022):

$$\min_{\mathbf{m} \in \{0,1\}^d} E_{\mathbf{u} \sim V} \left[\mathcal{D} \left(\Phi_c(\tilde{\mathbf{x}}), \Phi_c(\mathbf{x}) \right) \right] \quad \text{s.t. } \|\mathbf{m}\|_0 \leq s, \tag{1}$$

where $\mathcal{D} : [0,1]^c \times [0,1]^c \to \mathbb{R}_+$ quantifies the performance difference between the classifier's prediction for the original input $\mathbf{x}$ and the perturbed input $\tilde{\mathbf{x}}$, a term which is also referred to as the "distortion". $s \in \{1, ..., d\}$ controls the sparsity of the mask $\mathbf{m}$. The mean-squared error loss function is often used as the distortion measure $\mathcal{D}$ [2].

In practice, the RDE optimization framework in equation 1 is relaxed using both a continuous mask $\mathbf{m} \in [0,1]^d$ and a $\|\cdot\|_1$ norm:

$$\min_{\mathbf{m} \in [0,1]^d} E_{\mathbf{u} \sim V} \left[\mathcal{D} \left(\Phi_c(\tilde{\mathbf{x}}), \Phi_c(\mathbf{x}) \right) \right] + \lambda \|\mathbf{m}\|_1, \tag{2}$$

which can be optimized using gradient-based methods. Moreover, instead of applying the RDE framework to the original input domain, one can first transform $\mathbf{x}$ using an invertible function $\mathcal{F} : \mathbb{R}^d \to \mathbb{R}^k$, such as Wavelet transform (Kolek et al., 2022) and Shearlet transform (Kolek et al., 2023), and enforce sparsity in the frequency domain. In this scenario, the resulting objective becomes:

$$\min_{\mathbf{m} \in [0,1]^k} E_{\mathbf{u} \sim V} \left[\mathcal{D} \left(\Phi_c(\mathcal{F}^{-1}(\mathbf{m} \odot \mathcal{F}(\mathbf{x}) + (1-\mathbf{m}) \odot \mathbf{u})), \Phi_c(\mathbf{x}) \right) \right] + \lambda \|\mathbf{m}\|_1. \tag{3}$$

3.2 The Information Bottleneck (IB) Explanation Framework

An alternative, yet increasingly popular, objective for learning the mask $\mathbf{m}$ is based on the IB principle (Tishby et al., 1999; Gilad-Bachrach et al., 2003):

$$\min_{\mathbf{m} \in [0,1]^d} -I(\hat{y}; \tilde{\mathbf{x}}) + \beta I(\mathbf{x}; \tilde{\mathbf{x}}), \tag{4}$$

[1] Strictly speaking, these methods start with a constant value of 0.5 which is the expected value of an uniform distribution $\mathcal{U}(0,1)$.

[2] In fact, Kolek et al. (2022) report that the choice of the distortion function has limited effect on the performance of their proposed mask-based vision explanation methods.

where $I(\cdot;\cdot)$ represents the mutual information between two random variables, $\hat{y} = \Phi_c(\mathbf{x})$ is the black-box output, $\beta > 0$ is a Lagrange multiplier. Maximizing $I(\hat{y}; \tilde{\mathbf{x}})$ ensures that the perturbed input $\tilde{\mathbf{x}}$ retains sufficient information to approximate or explain the black-box output $\hat{y}$ (Bang et al., 2021), while minimizing $I(\mathbf{x}; \tilde{\mathbf{x}})$ encourages compression such that $\tilde{\mathbf{x}}$ does not capture any redundant information in $\mathbf{x}$ that is irrelevant to explain $\hat{y}$. In other words, the compression terms essentially encourages the conditional independence between $\tilde{\mathbf{x}}$ and $\mathbf{x}$, conditioning on $\hat{y}$ (Fischer, 2020).

Same to the RDE framework, the mask can be applied in a latent space (Schulz et al., 2020; Demir et al., 2021). In this case, the compression terms becomes $I(\mathbf{z}, \mathbf{z} \odot \mathbf{m} + (1 - \mathbf{m}) \odot \mathbf{u})$, in which $\mathbf{z} = f_l(\mathbf{x})$ refers to the l-th layer output of an instance $\mathbf{x}$.

Recently, the general idea of IB has been extended to the design of built-in interpretable deep learning architectures for image data (Choi et al., 2024) and graphs (Yu et al., 2021), where $\hat{y}$ is replaced by the ground-truth label y. However, our paper remains focused on post-hoc explanations.

From our perspective, there is a clear resemblance between equation 4 and the RDE objective in equation 2, as illustrated in Proposition 1 and Remark 1.

Proposition 1 *Maximizing $I(\hat{y}; \tilde{\mathbf{x}})$ is equivalent to minimizing the expected Kullback–Leibler (KL) divergence $\mathbb{E}\left(D_{KL}(\Phi_c(\tilde{\mathbf{x}}); \Phi_c(\mathbf{x}))\right)$, under the choice of KL divergence as the distortion measure in the RDE framework. Furthermore, the expected KL divergence provides an upper bound for the ℓ_1-loss, i.e., $\mathbb{E}\left(D_{KL}(\Phi_c(\tilde{\mathbf{x}}); \Phi_c(\mathbf{x}))\right) \geq \frac{1}{2\log 2}\mathbb{E}\left(\|\Phi_c(\tilde{\mathbf{x}}) - \Phi_c(\mathbf{x})\|_1^2\right)$.*

Remark 1 *In the IB implementation, the compression term $I(\mathbf{x}; \tilde{\mathbf{x}})$ can simply be replaced by an entropy term $H(\tilde{\mathbf{x}})$[3] (Strouse & Schwab, 2017; Kirsch et al., 2020). In fact, the simplest way to compress $H(\tilde{\mathbf{x}})$ is by encouraging the number of explanatory variables to be small, i.e., encouraging m to be sparse (Bang et al., 2021; Tao et al., 2020). This regularization can be expressed as $\|m\|_0$ or approximated with $\|m\|_1$.*

Proof. All proofs can be found in Appendix A.

4 METHOD

In this section, we present StartGrad, our novel gradient-based mask initialization algorithm. Unlike traditional approaches, StartGrad leverages gradient information to provide tailored initialization of masks based on the specific characteristic of the sample and model. This gradient-informed initialization can not only provide a more effective starting point at *initialization* in light of the previously discussed RDE and IB framework, but can also significantly boost the performance of subsequent mask-based explanation algorithms as we will show in section 5.

4.1 APPROXIMATING THE DISTORTION LOCALLY

Given a differentiable classifier Φ_c, we can use a first-order Taylor expansion to approximate the prediction for the distorted input $\tilde{\mathbf{x}}$:

$$\Phi_c(\tilde{\mathbf{x}}) \approx \Phi_c(\mathbf{x}) + \nabla \mathbf{x} \Phi_c(\mathbf{x}) \cdot \Delta\mathbf{x}, \tag{5}$$

where $\nabla_{\mathbf{x}}\Phi_c(\mathbf{x})$ represents the gradient of the model's prediction with respect to the original input feature $\mathbf{x}$ and $\Delta\mathbf{x} = \tilde{\mathbf{x}} - \mathbf{x}$ denotes the distortion induced due to the mask $\mathbf{m}$. To ensure that this approximation is accurate, we require that Φ_c is locally linear in a small neighborhood around $\mathbf{x}$. Formally, this neighborhood is defined as the open ball $B_\epsilon(\boldsymbol{x}) = \{\tilde{\boldsymbol{x}} \in \mathbb{R}^d : \|\tilde{\boldsymbol{x}} - \boldsymbol{x}\| < \epsilon\}$, where $\epsilon > 0$ ensures that higher-order terms in the Taylor expansion of Φ_c are negligible.

Using a $\|\cdot\|_p$ norm as a choice for the distortion function $\mathcal{D}$, we can leverage equation 5 to approximate the distortion:

$$\mathcal{D}\left(\Phi_c(\tilde{\mathbf{x}}), \Phi_c(\mathbf{x})\right) = \|\Phi_c(\tilde{\mathbf{x}}) - \Phi_c(\mathbf{x}))\|_p \approx \|\nabla_{\mathbf{x}}\Phi_c(\mathbf{x}) \cdot \Delta\mathbf{x}\|_p = \left(\sum_{i=1}^{d} |\nabla_{x_i}\Phi_c(\mathbf{x}) \cdot \Delta x_i|^p\right)^{\frac{1}{p}}. \tag{6}$$

[3] If the mapping from $\mathbf{x}$ to $\tilde{\mathbf{x}}$ is deterministic, we always have this equivalence; however, this may not hold true in practice, such as when unmasked regions of an image are substituted with random Gaussian noise.

From the approximation in equation 6, it follows that minimizing the distortion $\mathcal{D}\left(\Phi_c(\tilde{\mathbf{x}}), \Phi_c(\mathbf{x})\right)$ requires controlling the term $\|\nabla_{\mathbf{x}}\Phi_c(\mathbf{x}) \cdot \Delta\mathbf{x}\|_p$. In particular, we can leverage the gradient information to the minimize the distortion.[4] Features with small gradient magnitudes $|\nabla_{x_i}\Phi_c(\mathbf{x})|$ have less impact on the distortion term and can be therefore masked without significantly increasing $\mathcal{D}$. This insight leads to the following heuristic:

$$M_i = \begin{cases} 1 & \text{if } |\nabla_{x_i}\Phi_c(\mathbf{x})| \text{ is “large”}, \\ 0 & \text{if } |\nabla_{x_i}\Phi_c(\mathbf{x})| \text{ is “small”}. \end{cases} \tag{7}$$

Since $\nabla_{x_i}\Phi_c(\mathbf{x})$ indicates the sensitivity of the model's prediction to changes in the corresponding feature, masking features with "small" gradients ensures that any induced changes minimally impact the model prediction. This approach is crucial in maintaining the predictive performance while modifying or explaining the input, effectively balancing sparsity and distortion.

4.2 Gradient-based Mask initialization

A key challenge in our approach is setting appropriate thresholds to distinguish between "large" and "small" gradients when defining the mask in equation 7. As discussed in section 3.1, we avoid a binary threshold by leveraging continuous masks. Specifically, we define the absolute gradient of the model's prediction with respect to the class of interest as:

$$\mathcal{S}_c(\mathbf{x}) = \left|\frac{\partial\Phi_c(\mathbf{x})}{\partial\mathbf{x}}\right|. \tag{8}$$

For multi-channel inputs such as RGB images, we take the maximum value across channels. We then apply a transformation function $\mathcal{T} : \mathbb{R}_+^d \to [0,1]^d$, which normalizes the gradient signal $\mathcal{S}_c(\mathbf{x})$ into the range $[0,1]^d$. The primary challenge lies in selecting a suitable transformation function that preserves the relative importance of each feature.

Empirically, we observe that the absolute gradient values follow a highly skewed distribution, similar to a Laplace distribution[5] (see Appendix D). Given this skewness, we employ the Quantile Transformation Function (QTF), which maps the gradient values to a uniform distribution $\mathcal{U}(0,1)$. The QTF has two key advantages: (1) it handles the skewness in the data effectively, and (2) it preserves the relative feature importance based on the absolute gradient values. Furthermore, using the QTF facilitates direct comparison with the standard uniform mask initialization scheme, thereby simplifying the evaluation of our proposed framework. Algorithm 1 gives pseudocode for our gradient-based initialization method which we call StartGrad.[6]

Algorithm 1: Gradient-based Mask Initialization (StartGrad)

```
Input          : Pre-trained classifier $\Phi_c$, input samples $\{\boldsymbol{x}_i\}_{i=1}^N$ with $\boldsymbol{x}_i \in \mathbb{R}^d$, quantile
                 transformation $\mathcal{T} : \mathbb{R}_+^d \to [0,1]^d$
Hyperparameters: Output distribution $\mathcal{U}(0,1)$, number of quantiles (bins) for $\mathcal{T}$
Output         : Masks $\mathbf{M} \in [0,1]^{N \times d}$
Initialize mask list $\mathbf{M} := []$ and quantile transformer $\mathcal{T}$;
for $i \leftarrow 1$ to $N$ do
    $c_i \leftarrow \arg\max(\Phi_c(\boldsymbol{x}_i))$ // Class prediction
    $\mathbf{S}_i \leftarrow |\nabla_{\boldsymbol{x}_i}\Phi_c(\boldsymbol{x}_i)|_c$ // Gradient magnitudes
    $\mathbf{m}_i \leftarrow \mathcal{T}(\mathbf{S}_i)$ // Quantile transform
    Append $\mathbf{m}_i$ to $\mathbf{M}$;
end for
return $\mathbf{M}$
```

[4] An alternative strategy is to trivially set $\Delta\mathbf{x} = \mathbf{0}_d$, i.e. $\mathbf{m} = \mathbf{1}_d$, thereby keeping all features and eliminating distortion entirely. However, this solution fails to satisfy the sparsity objective of the RDE framework and is thus suboptimal for the overall task.

[5] This behavior has been previously observed in other domains as well, such as time series (Ismail et al., 2020) and graph-based models (Xie et al., 2022)

[6] For simplicity, we assume for the pseudocode in algorithm 1 that we operate in the same input space $\mathcal{X}$. In Appendix B, a more general version is provided where we we use an invertible and differentiable function $\mathcal{F}$. A visualization of StartGrad is also provided in the Appendix, as it was moved there due to space constraints.

4.3 Balancing sparsity and distortion under the RDE framework

As outlined in section 3.1, the primary objective of mask-based explanation algorithms is typically to identify a sparse mask that minimally distorts the model's predictions.[7] While our previous discussion in 4.1 has centered on minimizing distortion at initialization, we now turn our attention to sparsity, comparing the standard initialization strategies of all-ones and uniform to our newly introduced algorithm StartGrad. Specifically, we provide formal proofs that StartGrad is *provably superior at initialization* in balancing the trade-off between distortion and sparsity, as formalized in the RDE framework, when compared to other mask initialization strategies. To the best of our knowledge, this represents the first formal comparison of different mask initialization methods.

Proposition 2 *Let $\Phi_c(\boldsymbol{x})$ with $\Phi_c : \mathbb{R}^d \to [0,1]^c$ be a classifier that is differentiable and locally linear in the neighborhood of an input $\mathbf{x} \in \mathbb{R}^d$. Formally, this neighborhood is defined as the open ball $B_\epsilon(\boldsymbol{x}) = \{\tilde{\boldsymbol{x}} \in \mathbb{R}^d : \|\tilde{\boldsymbol{x}} - \boldsymbol{x}\| < \epsilon\}$, where $\epsilon > 0$ ensures that higher-order terms in the Taylor expansion of Φ_c are negligible. Additionally, let $\mathbf{m}_{unif}$ represent the uniformly-initialized mask and $\mathbf{m}_{grad}$ the mask initialized based on the gradient of $\Phi_c(\mathbf{x})$ with respect to $\mathbf{x}$, transformed using a quantile function so that $\mathbf{m}_{grad} \in [0,1]^d$. Then, given a norm $\|\cdot\|_p$ with $p \geq 1$ as a choice for the distortion function $\mathcal{D}$, the following inequality holds at initialization:*

$$\mathbb{E}_{\mathbf{m}_{grad},\mathbf{u}\sim V}\left[D\left(\Phi_c(\tilde{\mathbf{x}}_{grad}), \Phi_c(\mathbf{x})\right) + \lambda\|\mathbf{m}_{grad}\|_1\right] \leq \mathbb{E}_{\mathbf{m}_{unif},\mathbf{u}\sim V}\left[D\left(\Phi_c(\tilde{\mathbf{x}}_{unif}), \Phi_c(\mathbf{x})\right) + \lambda\|\mathbf{m}_{unif}\|_1\right]$$

with $\tilde{\mathbf{x}}_{unif} = \mathbf{m}_{unif} \odot \mathbf{x} + (1 - \mathbf{m}_{unif}) \odot \mathbf{u}$ and $\tilde{\mathbf{x}}_{grad} = \mathbf{m}_{grad} \odot \mathbf{x} + (1 - \mathbf{m}_{grad}) \odot \mathbf{u}$, where $\mathbf{u} \in \mathbb{R}^d$ is a random perturbation drawn from a predefined distribution V and λ is a hyperparameter encouraging sparsity in the masks.

Proposition 3 *Let $\Phi_c(\boldsymbol{x})$ with $\Phi_c : \mathbb{R}^d \to [0,1]^c$ be a classifier and let $\boldsymbol{m}_{ones}$ denote the mask initialized with all ones, i.e. $\boldsymbol{m}_{ones} = \mathbf{1}_d$ where $\mathbf{1}_d$ is a vector of ones with dimension d and $\mathbf{m}_{grad}$ denote the mask initialized based on the gradient of $\Phi_c(\mathbf{x})$ with respect to $\mathbf{x}$, transformed using a quantile function so that $\mathbf{m}_{grad} \in [0,1]^d$. Given a norm $\|\cdot\|_p$ with $p \geq 1$ as a choice for the distortion function $\mathcal{D}$, the following inequality holds at initialization:*

$$\mathbb{E}_{\mathbf{m}_{grad},\mathbf{u}\sim V}\left[D\left(\Phi_c(\tilde{\mathbf{x}}_{grad}), \Phi_c(\mathbf{x})\right) + \lambda\|\mathbf{m}_{grad}\|_1\right] \leq \mathbb{E}_{\mathbf{m}_{ones},\mathbf{u}\sim V}\left[D\left(\Phi_c(\tilde{\mathbf{x}}_{ones}), \Phi_c(\mathbf{x})\right) + \lambda\|\mathbf{m}_{ones}\|_1\right]$$

for

$$\frac{2\left(1 + (2-1)\mathbb{1}_{c\geq 2}\right)^{\frac{1}{p}}}{d} \leq \lambda$$

with $\tilde{\mathbf{x}}_{ones} = \mathbf{m}_{ones} \odot \mathbf{x} + (1 - \mathbf{m}_{ones}) \odot \mathbf{u}$ and $\tilde{\mathbf{x}}_{grad} = \mathbf{m}_{grad} \odot \mathbf{x} + (1 - \mathbf{m}_{grad}) \odot \mathbf{u}$, where $\mathbf{u} \in \mathbb{R}^d$ is a random perturbation drawn from a predefined distribution V and λ is a hyperparameter encouraging sparsity in the masks.

Proof. All proofs can be found in Appendix A.

Remark 2 *In case we want to explain the class prediction of a classifier $\Phi_c(\boldsymbol{x})$ for a specific category, i.e. $c = 1$, Proposition 3 reduces to*

$$\frac{2^{\frac{1}{p}}}{d} \leq \lambda$$

Remark 3 *Proposition 3 holds in the same manner for the comparison between $\mathbf{m}_{unif}$ and $\mathbf{m}_{ones}$.*

Proposition 2 and 3 together provide a principled approach for selecting an initialization strategy within the RDE framework. Our proposed gradient-based initialization algorithm yields an expected loss in terms of distortion and sparsity that is less than or equal to that of uniform initialization. Additionally, StartGrad outperforms all-ones initialization in high-dimensional settings (where d is large) which is a typical setting for deep learning models, and consequently, mask-based explanation approaches.

[7] In addition to sparsity and distortion, some mask-based algorithms also regularize the smoothness of the mask (Fong & Vedaldi, 2017). Furthermore, rather than minimizing distortion, certain algorithms aim to find the smallest mask that induces the maximum distortion (Enguehard, 2023).

5 EXPERIMENTS

In this section, we evaluate our gradient-based initialization scheme using state-of-the-art mask-based explanation methods across two challenging data modalities: vision and time series.[8] Vision is a domain where mask-based explanation methods are well-established, while time-series data, though less studied in the XAI community, is critical in fields like healthcare and finance, gaining more attention in recent years (Crabbé & Van Der Schaar, 2021; Enguehard, 2023; Liu et al., 2024; Queen et al., 2023; Zichuan Liu, 2024). These domains present distinct challenges for gradient estimation—vision involves spatial complexity, while time-series data introduces temporal dependencies (Ismail et al., 2020). This cross-domain evaluation demonstrates how StartGrad adapts to varied data characteristics and model architectures, underscoring the robustness and potential of our proposed initialization algorithm.

5.1 VISION

We evaluate StartGrad using two state-of-the-art mask-based explanation methods: WaveletX, ShearletX (Kolek et al., 2023) and a well-established third method, i.e. PixelMask (Fong & Vedaldi, 2017). These vision methods were chosen as they apply masking in different input spaces and vary in their objective formulations, creating a diverse and challenging testbed for our proposed initialization method.

For quantitative evaluation, we use the conciseness-preciseness (CP) Pixel and L1 scores introduced by Kolek et al. (2023), calculated on 500 random samples from the ImageNet validation dataset (Deng et al., 2009). These information-theoretic metrics are designed to reward masks that preserve classification accuracy while minimizing the amount of information used, making them well-suited for mask-based methods that prioritize sparsity and compressed explanations. Unlike faithfulness metrics (e.g., insertion and deletion), which rely on an inherent ranking of feature importance, CP scores are specifically designed for methods like ours that optimize binary masks without ordering feature relevance (Kolek et al., 2023). Additionally, using CP scores ensures consistency and comparability with the baseline methods in Kolek et al. (2023).[9]

As in Kolek et al. (2023), we use a pretrained ResNet18 model (He et al., 2016) and VGG16 (Simonyan & Zisserman, 2014) for classification. We also evaluate StartGrad on vision transformers (Dosovitskiy et al., 2021) and swin transformers (Liu et al., 2022). Due to space constraints, only results for the ResNet18 model are shown in the main body of the paper, with results for other architectures provided in Appendix F.1, F.4, and F.3. The qualitative findings discussed here are vastly consistent across all tested models. For all mask-based vision models, we use the default parameter settings as reported in the respective papers to facilitate comparisons. Further details on the explanation models and hyperparameters are in Appendix C.1. Additional results and ablation studies are provided in section D.3 and F. In Appendix F.18, F.19, F.20 we also provide visual results.

Results. Table 1 shows the median performance difference between StartGrad and both all-ones and uniform initialization strategies. Across all mask-based explanation methods, StartGrad significantly improves performance compared to the uniform initialization which shares the underlying initialization distribution. This underscores the advantage of using gradient signals to guide the mask initialization process.

For PixelMask (Fong & Vedaldi, 2017) and WaveletX (Kolek et al., 2023), StartGrad particularly enhances performance at early iteration steps, whereas at the final iteration step (300) no statistical differences can be observed for these two mask-based explanation methods. However, initializing ShearletX (Kolek et al., 2023) with our proposed StartGrad algorithm leads to a substantial and statistically significant performance improvement at all iteration steps.

In addition to measuring performance at a given iteration step, we examine how StartGrad reduces runtime (as measured in iteration steps) when aiming for a specific target metric. However, defining a universally "good" target score is inherently challenging due to the absence of a clear threshold. To tackle this, we define the per image target score as the highest attainable score across all three initializations and measure subsequently the iterations required to this specific normalized score.

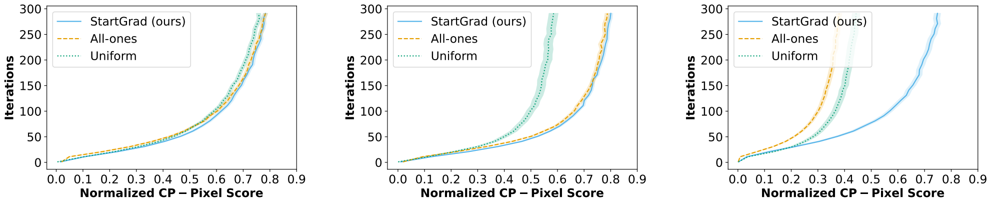
[8] Our code is publicly available at `https://github.com/BuelentUendes/StartGrad`

[9] For a more detailed justification of using the CP scores in place of faithfulness, see Appendix C.2.2.

Table 1: Median pairwise performance difference between StartGrad and alternative initialization methods across iteration steps for 500 ImageNet (Deng et al., 2009) samples using a ResNet18 (He et al., 2016) classifier. Baseline refers to the initialization scheme originally used for each mask explanation method. Statistical significance is marked as ***, **, and * for 1%, 5%, and 10% levels, respectively, using a one-sided Wilcoxon signed-rank test with Bonferroni correction. For PixelMask (Fong & Vedaldi, 2017), CP-Pixel and CP-L1 scores are identical as it operates only in the pixel space.

		△ CP-Pixel ↑			△ CP-L1 ↑		
		Iteration steps			Iteration steps		
Method	Reference initialization	50	100	300	50	100	300
PixelMask	All-ones (baseline)	2.09^{***}	0.72	0.33	2.09^{***}	0.72	0.33
	Uniform	1.11^{***}	1.26^{***}	1.12^{***}	1.11^{***}	1.26^{***}	1.12^{***}
WaveletX	All-ones (baseline)	2.55^{***}	0.73^{**}	0.11	0.71^{***}	0.37^{***}	0.14
	Uniform	2.35^{***}	3.39^{***}	5.20^{***}	0.60^{***}	0.88^{***}	1.04^{***}
ShearletX	All-ones (baseline)	52.91^{***}	135.28^{***}	236.37^{***}	2.89^{***}	20.67^{***}	51.82^{***}
	Uniform	58.65^{***}	130.71^{***}	208.74^{***}	3.09^{***}	33.42^{***}	65.57^{***}

As shown in Figure 1, StartGrad outperforms the other two initializations strategies for PixelMask (Fong & Vedaldi, 2017) and WaveletX (Kolek et al., 2023) for medium target scores and for ShearletX (Kolek et al., 2023) beginning from a score of 0.3.



(a) PixelMask (b) WaveletX (c) ShearletX

Figure 1: Normalized CP-Pixel score vs. iteration steps for the PixelMask (Fong & Vedaldi, 2017) (a), WaveletX (Kolek et al., 2023) (b), and ShearletX (Kolek et al., 2023) (c) models, comparing three initialization methods: StartGrad (ours) (solid line), All-ones (dashed line), and Uniform (dotted line). The curves represent the average number of iteration steps needed to reach a normalized target CP-Pixel score. This target score is defined as the highest CP-Pixel value achieved across all three initialization methods for each model, then normalized by the overall maximum score observed. The shaded regions represent the standard error across 500 randomly selected ImageNet (Deng et al., 2009) validation samples, evaluated using a pretrained ResNet18 (He et al., 2016) classifier.

Given the strong performance of the StartGrad initialization method for the ShearletX (Kolek et al., 2023) model, we seek to evaluate how quickly a mask-based algorithm initialized with StartGrad can reach the final performance levels of the all-ones and uniform initialization strategies. As it can be seen from table 2, we achieve large speedups up to 6.73, further illustrating the potential of StartGrad and designing mask-based initialization strategies in general. To further understand the dynamics of the optimization process and the effectiveness of StartGrad, we analyze its role in guiding mask-based explanation methods. As shown in Appendix F.6, F.7, F.8, initializing the mask using gradient signals, as implemented in StartGrad, provides a crucial early advantage by guiding the optimization process toward regions of minimal distortion, consistent with theoretical predictions. This early advantage is critical, as uniform initialization fails to catch up in terms of distortion reduction throughout the optimization process which underlines the importance of an effective initialization.

Table 2: Average iteration steps for the ShearletX (Kolek et al., 2023) method initialized with StartGrad to match the performance of the all-ones or uniform initializations. Reported are the iteration steps to match the interquartile mean (IQM) and median of the CP-Pixel and CP-L1 scores, along with speedup calculated as the ratio of reference initialization to StartGrad iteration steps. All experiments use 500 randomly selected ImageNet (Deng et al., 2009) samples with a pretrained ResNet18 (He et al., 2016) classifier, and metrics are computed over 300 iterations. Average and standard errors are estimated via bootstrapping (250 resamples). For each metric, ↑ denotes higher is better and ↓ lower is better. Values that are two standard errors away from the baseline (300 iterations, 1 for speedup) are highlighted in bold.

Reference	Target metric (at 300 iterations)	CP-Pixel		CP-L1	
		Iterations ↓	Speedup ↑	Iterations ↓	Speedup ↑
All-ones (baseline)	IQM	**44.57 ± 2.36**	**6.73 ± 0.35**	**71.09 ± 6.47**	**4.24 ± 0.38**
	Median	**49.60 ± 3.39**	**6.06 ± 0.41**	**87.70 ± 9.98**	**3.45 ± 0.38**
Uniform	IQM	**45.50 ± 3.18**	**6.60 ± 0.45**	**72.92 ± 8.68**	**4.16 ± 0.47**
	Median	**50.58 ± 4.26**	**5.95 ± 0.49**	**94.44 ± 15.10**	**3.24 ± 0.47**

5.2 Time Series

Following Tonekaboni et al. (2020), Crabbé & Van Der Schaar (2021) and Enguehard (2023), we test StartGrad on two commonly used synthetic benchmark datasets, i.e. state and switch-feature data both of which use a hidden Markov model (HMM) to generate the data. In line with Crabbé & Van Der Schaar (2021); Enguehard (2023) we train a one-layer GRU (Cho et al., 2014) for each of the experiments. As a baseline mask-based explanation method we use the recently introduced ExtremalMask (Enguehard, 2023), which showed state-of-the-art performance on the aforementioned datasets. As a perturbation model, we use a bidirectional one-layer GRU (Cho et al., 2014) model. We evaluate StartGrad using the following four metrics: area under recall (AUR), area under precision (AUP), information (I) and mask entropy (E), where the last two were introduced by Crabbé & Van Der Schaar (2021). For both synthetic datasets, we use the same experimental setup as in previous studies (Crabbé & Van Der Schaar, 2021; Enguehard, 2023; Liu et al., 2024), i.e. we generate 1,000 samples and train the classifier on 800 training examples, and evaluate the performance on the remaining 200 samples while reporting results across five folds. We use the default hyperparameter settings as well as the same random seed as used by Enguehard (2023) to facilitate fair comparisons.

A complete description of the datasets, models and hyperparameters used in this study can be found in the Appendix C.3. Due to space constraints, we only report the results on the dwitch-feature dataset for the preservation game version of the ExtremalMask (Enguehard, 2023) model. All remaining results are provided in section E.

Results. As illustrated in Table 3, initializing ExtremalMask (Enguehard, 2023) with StartGrad yields strong performance improvements especially in the early iteration steps. At iteration 50, StartGrad outperforms the baseline (uniform) in all metrics, with notably performance improvements. Compared to the all-ones initialization, StartGrad provides performance improvements in 3 out 4 metrics, with the exception of AUP which is slightly lower. However, the all-ones mask initialization has a low corresponding AUR score, indicating that ExtremalMask (Enguehard, 2023) initialized with all-ones method misses a lot of important features. This advantage is maintained through iteration 100. By iteration 500, all methods converge to similar final performance, demonstrating that StartGrad provides a strong initial boost without compromising the overall outcome for ExtremalMask (Enguehard, 2023). The findings are consistent across datasets and objective formulations as shown in Appendix E.

6 Limitations and future research

Despite the promising results of the StartGrad initialization algorithm, few limitations require attention:

Table 3: Average performance for the ExtremalMask (Enguehard, 2023) explanation method (preservation game objective) across iteration steps for the switch-feature dataset when initialized with StartGrad, all-ones and uniformly. The reported numbers are the mean and standard deviation across five folds. For each metric, ↑ indicates that higher is better, and ↓ that lower is better. Baseline refers to the initialization scheme originally used for the respective mask explanation method. Outcomes that are one standard deviation away from the second-best ones are highlighted in bold.

		Iteration steps			
Metric	Initialization	50	100	300	500
AUP ↑	Uniform (baseline)	0.849 ± 0.049	0.942 ± 0.039	0.982 ± 0.008	0.986 ± 0.004
	All-ones	**0.952** ± **0.003**	0.967 ± 0.004	0.986 ± 0.003	0.986 ± 0.003
	StartGrad (ours)	0.922 ± 0.004	**0.984** ± **0.003**	0.987 ± 0.003	0.987 ± 0.003
AUR ↑	Uniform (baseline)	0.746 ± 0.029	0.737 ± 0.023	0.747 ± 0.020	0.748 ± 0.020
	All-ones	0.696 ± 0.015	**0.769** ± **0.017**	0.751 ± 0.020	0.750 ± 0.020
	StartGrad (ours)	**0.805** ± **0.014**	**0.758** ± **0.020**	0.751 ± 0.020	0.751 ± 0.021
I [10^5] ↑	Uniform (baseline)	1.069 ± 0.085	1.876 ± 0.229	2.513 ± 0.079	2.533 ± 0.078
	All-ones	1.905 ± 0.061	2.446 ± 0.091	2.541 ± 0.079	2.539 ± 0.072
	StartGrad (ours)	**2.498** ± **0.105**	**2.533** ± **0.083**	2.544 ± 0.082	2.540 ± 0.082
E [10^4] ↓	Uniform (baseline)	2.325 ± 0.151	1.659 ± 0.197	1.281 ± 0.071	1.267 ± 0.072
	All-ones	1.984 ± 0.066	1.582 ± 0.085	1.277 ± 0.070	1.276 ± 0.072
	StartGrad (ours)	**1.688** ± **0.103**	**1.370** ± **0.074**	1.276 ± 0.069	1.278 ± 0.069

First, StartGrad's effectiveness relies on accurate gradient estimation, making it potentially sensitive to noise. In Appendix F.5, we assess the impact of noisy gradients in the vision domain by adding Gaussian noise to the gradient signal. Appendix F.13 examines uninformative gradients by shuffling values and explores an adversarial setting where gradient-mask associations are inverted. Finally, Appendix F.10 evaluates sensitivity to gradient estimation methods using SmoothGrad (Smilkov et al., 2017). Our findings show StartGrad remains robust to noise and alternative estimation techniques, matching uniform initialization in uninformative scenarios. However, its failure to recover from adversarially initialized masks highlights the importance of proper initialization.

Second, the choice of the transformation function $\mathcal{T}$ that maps the gradients into a continuous mask is crucial for StartGrad as illustrated in Appendix F.11. Future work could investigate alternative transformations functions other than the QTF. Another promising direction for future work is filtering high-frequency artifacts in gradient signals before applying the QTF transform, as this has been shown to improve the quality of the resulting gradient signal (Muzellec et al., 2024).

Third, compared to other initialization techniques, StartGrad involves some computational overhead. Yet, as illustrated in Appendix D.2, the additional computational costs are negligible, especially in light of the presented performance gains and speedups.

7 Conclusion

In this work, we present the first comprehensive theoretical and empirical analysis of mask initialization techniques and introduce StartGrad, a novel gradient-based mask initialization method. Specifically designed for post-hoc mask-based explanation methods, we demonstrate that StartGrad is provably superior in terms of distortion and sparsity at initialization compared to other techniques. Despite its simplicity, our experiments demonstrate that StartGrad enhances the optimization process of various state-of-the-art mask-based explanation methods, particularly by improving performance early on.

Given its effectiveness, StartGrad can serve as a strong baseline initialization method moving forward. We also hope that this work sparks further interest in exploring strategies for effective initialization techniques in mask-based explanation methods in an attempt to improve their practicality in real-world applications.

ACKNOWLEDGMENTS

This work is funded by Stress in Action. The research project 'Stress in Action': www.stress-in-action.nl is financially supported by the Dutch Research Council and the Dutch Ministry of Education, Culture and Science (NWO gravitation grant number 024.005.010).

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 9525–9536, Red Hook, NY, USA, 2018. Curran Associates Inc.

Seojin Bang, Pengtao Xie, Heewook Lee, Wei Wu, and Eric Xing. Explaining a black-box by using a deep variational information bottleneck approach. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11396–11404, 2021.

Wouter Bant, Ádám Divák, Jasper Eppink, and Floris Six Dijkstra. On the reproducibility of: "learning perturbations to explain time series predictions". *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=nPZgtpfgIx`. Reproducibility Certification.

Clément L Canonne. A short note on an inequality between kl and tv. *arXiv preprint arXiv:2202.07198*, 2022.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches, 2014. URL `https://arxiv.org/abs/1409.1259`.

Changkyu Choi, Shujian Yu, Michael Kampffmeyer, Arnt-Børre Salberg, Nils Olav Handegard, and Robert Jenssen. Dib-x: Formulating explainability principles for a self-explainable model through information theoretic learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7170–7174, 2024.

Jonathan Crabbé and Mihaela Van Der Schaar. Explaining time series predictions with dynamic masks. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2166–2177. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/crabbe21a.html`.

Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6970–6979, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Ugur Demir, Ismail Irmakci, Elif Keles, Ahmet Topcu, Ziyue Xu, Concetto Spampinato, Sachin Jambawalikar, Evrim Turkbey, Baris Turkbey, and Ulas Bagci. Information bottleneck attribution for visual explanations of diagnosis and prognosis. In *Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12*, pp. 396–405. Springer, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

Joseph Enguehard. Learning perturbations to explain time series predictions. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 9329–9342. PMLR, 2023. URL https://proceedings.mlr.press/v202/enguehard23a.html. ISSN: 2640-3498.

Ian Fischer. The conditional entropy bottleneck. *Entropy*, 22(9):999, 2020.

Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3449–3457, 2017. doi: 10.1109/ICCV.2017.371.

Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3681–3688, Jul. 2019. doi: 10.1609/aaai.v33i01.33013681. URL https://ojs.aaai.org/index.php/AAAI/article/view/4252.

Ran Gilad-Bachrach, Amir Navot, and Naftali Tishby. An information theoretic tradeoff between complexity and accuracy. In Bernhard Schölkopf and Manfred K. Warmuth (eds.), *Learning Theory and Kernel Machines*, volume 2777, pp. 595–609. Springer Berlin Heidelberg, 2003. ISBN 978-3-540-40720-1 978-3-540-45167-9. doi: 10.1007/978-3-540-45167-9_43. URL http://link.springer.com/10.1007/978-3-540-45167-9_43. Series Title: Lecture Notes in Computer Science.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington (eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL https://proceedings.mlr.press/v9/glorot10a.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Aya Abdelsalam Ismail, Mohamed Gunady, Hector Corrada Bravo, and Soheil Feizi. Benchmarking deep learning interpretability in time series predictions. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6441–6452. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/47a3893cc405396a5c30d91320572d6d-Abstract.html.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL https://api.semanticscholar.org/CorpusID:6628106.

Andreas Kirsch, Clare Lyle, and Yarin Gal. Unpacking information bottlenecks: Unifying information-theoretic objectives in deep learning. *arXiv preprint arXiv:2003.12537*, 2020.

Stefan Kolek, Duc Anh Nguyen, Ron Levie, Joan Bruna, and Gitta Kutyniok. Cartoon explanations of image classifiers. In *European Conference of Computer Vision (ECCV)*, 2022.

Stefan Kolek, Robert Windesheim, Hector Andrade Loarca, Gitta Kutyniok, and Ron Levie. Explaining image classifiers with multiscale directional image representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023.

Gitta Kutyniok, Wang-Q Lim, and Rafael Reisenhofer. Shearlab 3d: Faithful digital shearlet transforms based on compactly supported shearlets. *ACM Trans. Math. Softw.*, 42(1), January 2016. ISSN 0098-3500. doi: 10.1145/2740960. URL https://doi.org/10.1145/2740960.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12009–12019, June 2022.

Zichuan Liu, Yingying Zhang, Tianchun Wang, Zefan Wang, Dongsheng Luo, Mengnan Du, Min Wu, Yi Wang, Chunlin Chen, Lunting Fan, and Qingsong Wen. Explaining time series via contrastive and locally sparse perturbations. In *Proceedings of the 12th International Conference on Learning Representations*, pp. 1–21, 2024.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Jan Macdonald, Stephan Wäldchen, Sascha Hauch, and Gitta Kutyniok. A rate-distortion framework for explaining neural network decisions, 2019.

Sabine Muzellec, Thomas FEL, Victor Boutin, Léo Andéol, Rufin VanRullen, and Thomas Serre. Saliency strikes back: How filtering out high frequencies improves white-box explanations. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=eC1OOpOGZW.

Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

Mark S Pinsker. Information and information stability of random variables and processes. *Holden-Day*, 1964.

Zoe Piran, Ravid Shwartz-Ziv, and Naftali Tishby. The dual information bottleneck. *arXiv preprint arXiv:2006.04641*, 2020.

Owen Queen, Thomas Hartvigsen, Teddy Koker, He Huan, Theodoros Tsiligkaridis, and Marinka Zitnik. Encoding time-series explanations through self-supervised model behavior consistency. In *Proceedings of Neural Information Processing Systems, NeurIPS*, 2023.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ”why should i trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, pp. 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. Restricting the flow: Information bottlenecks for attribution. In *International Conference on Learning Representations*, 2020.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017. doi: 10.1109/ICCV.2017.74.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. URL https://api.semanticscholar.org/CorpusID:14124313.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*, 2014.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES ’20, pp. 180–186, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371100. doi: 10.1145/3375627.3375830. URL https://doi.org/10.1145/3375627.3375830.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. In *Workshop on Visualization for Deep Learning, ICML*, 2017.

DJ Strouse and David J Schwab. The deterministic information bottleneck. *Neural computation*, 29 (6):1611–1630, 2017.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 06–11 Aug 2017. URL `https://proceedings.mlr.press/v70/sundararajan17a.html`.

Ruo Yu Tao, Vincent François-Lavet, and Joelle Pineau. Novelty search in representational space for sample efficient exploration. *Advances in Neural Information Processing Systems*, 33:8114–8126, 2020.

MTCAJ Thomas and A Thomas Joy. *Elements of information theory*. Wiley-Interscience, 2006.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pp. 368–377, 1999.

Sana Tonekaboni, Shalmali Joshi, Kieran Campbell, David K Duvenaud, and Anna Goldenberg. What went wrong and when? instance-wise feature importance for time-series black-box models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 799–809. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/08fa43588c2571ade19bc0fa5936e028-Paper.pdf`.

Alfredo Vellido. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Comput. Appl.*, 32(24):18069–18083, December 2020. ISSN 0941-0643. doi: 10.1007/s00521-019-04051-w. URL `https://doi.org/10.1007/s00521-019-04051-w`.

Yaochen Xie, Sumeet Katariya, Xianfeng Tang, Edward Huang, Nikhil Rao, Karthik Subbian, and Shuiwang Ji. Task-agnostic graph explanations. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 12027–12039. Curran Associates, Inc., 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/4eb7f0abf16d08e50ed42beb1e22e782-Paper-Conference.pdf`.

Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. *GNNExplainer: generating explanations for graph neural networks*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Junchi Yu, Tingyang Xu, Yu Rong, Yatao Bian, Junzhou Huang, and Ran He. Graph information bottleneck for subgraph recognition. In *International Conference on Learning Representations*, 2021.

Jimeng Shi Xu Zheng Zhuomin Chen Lei Song Wenqian Dong Jayantha Obeysekera Farhad Shirani Dongsheng Luo Zichuan Liu, Tianchun Wang. Timex++: Learning time-series explanations with information bottleneck. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

A Proofs

A.1 Resemblance between IB objective and RDE objective

The IB optimization trade-off can be considered as a generalized rate-distortion problem (Tishby et al., 1999; Piran et al., 2020) with the distortion function between an instance $\mathbf{x}$ and a (compressed) representation $\tilde{\mathbf{x}}$ taken as the KL-divergence between their predictions of the black-box output $\hat{y}$:

$$\begin{aligned} d_{\text{IB}}(\mathbf{x};\tilde{\mathbf{x}}) &= D_{\text{KL}}[p(\hat{y}|\mathbf{x});p(\hat{y}|\tilde{\mathbf{x}})] \\ &= \sum_{\hat{y}} p(\hat{y}|\mathbf{x}) \log\left(\frac{p(\hat{y}|\mathbf{x})}{p(\hat{y}|\tilde{\mathbf{x}})}\right). \end{aligned} \tag{9}$$

The expected distortion $\mathbb{E}[d_{\text{IB}}(\mathbf{x};\tilde{\mathbf{x}})]$ simply reduces to $I(\mathbf{x};\hat{y}) - I(\tilde{\mathbf{x}};\hat{y})$. This is because:

$$\begin{aligned} \mathbb{E}_{p(\mathbf{x},\tilde{\mathbf{x}})}[d_{\text{IB}}(\mathbf{x};\tilde{\mathbf{x}})] &= \mathbb{E}[\sum_{\hat{y}} p(\hat{y}|\mathbf{x}) \log p(\hat{y}|\mathbf{x})] - \mathbb{E}[\sum_{\hat{y}} p(\hat{y}|\mathbf{x}) \log p(\hat{y}|\tilde{\mathbf{x}})] \\ &= \sum_{\mathbf{x}}\sum_{\tilde{\mathbf{x}}}\sum_{\hat{y}} p(\mathbf{x},\tilde{\mathbf{x}})p(\hat{y}|\mathbf{x}) \log p(\hat{y}|\mathbf{x}) - \sum_{\mathbf{x}}\sum_{\tilde{\mathbf{x}}}\sum_{\hat{y}} p(\mathbf{x},\tilde{\mathbf{x}})p(\hat{y}|\mathbf{x}) \log p(\hat{y}|\tilde{\mathbf{x}}) \\ &= \sum_{\mathbf{x}}\sum_{\tilde{\mathbf{x}}}\sum_{\hat{y}} p(\mathbf{x},\tilde{\mathbf{x}},\hat{y}) \log p(\hat{y}|\mathbf{x}) - \sum_{\mathbf{x}}\sum_{\tilde{\mathbf{x}}}\sum_{\hat{y}} p(\mathbf{x},\tilde{\mathbf{x}},\hat{y}) \log p(\hat{y}|\tilde{\mathbf{x}}) \\ &= \mathbb{E}[\log p(\hat{y}|\mathbf{x})] - \mathbb{E}[\log p(\hat{y}|\tilde{\mathbf{x}})] \\ &= -H(\hat{y}|\mathbf{x}) + H(\hat{y}) + H(\hat{y}|\tilde{\mathbf{x}}) - H(\hat{y}) \\ &= I(\hat{y};\mathbf{x}) - I(\hat{y};\tilde{\mathbf{x}}), \end{aligned} \tag{10}$$

where the third line of equation uses the Markov chain property $\hat{y} \leftarrow \mathbf{x} \rightarrow \tilde{\mathbf{x}}$, i.e., $p(\hat{y}|\mathbf{x},\tilde{\mathbf{x}}) = p(\hat{y}|\mathbf{x})$. Line five uses the definition of conditional entropy, i.e. $H(\hat{y}|x) = -\mathbb{E}[\log p(\hat{y}|x)]$. and line six the definition of mutual information $I(\hat{y};\boldsymbol{x}) = H(\hat{y}) - H(\hat{y}|\boldsymbol{x})$ respectively.

Since $I(\hat{y};\mathbf{x})$ is a constant that is independent of the optimization, minimizing the expected IB distortion $\mathbb{E}[D_{\text{KL}}[p(\hat{y}|\mathbf{x});p(\hat{y}|\tilde{\mathbf{x}})]]$ is equivalent to maximizing $I(\hat{y};\tilde{\mathbf{x}})$.

Note that, in neural networks implementation, $p(\hat{y}|\mathbf{x})$ is approximated by the output of last softmax layer, i.e., $\Phi_c(\mathbf{x})$. Similarly, $p(\hat{y}|\tilde{\mathbf{x}})$ is approximated by $\Phi_c(\tilde{\mathbf{x}})$.

On the other hand, due to the Pinsker's inequality (Pinsker, 1964; Canonne, 2022) which implies that for any two distributions p and q,

$$\min\{D_{\text{KL}}(p;q), D_{\text{KL}}(q;p)\} \geq \frac{1}{2\log 2}\|p-q\|_1^2, \tag{11}$$

from which we obtain $D_{\text{KL}}(\Phi_c(\tilde{\mathbf{x}});\Phi_c(\mathbf{x})) \geq \frac{1}{2\log 2}\|\Phi_c(\tilde{\mathbf{x}}) - \Phi_c(\mathbf{x})\|_1^2$.

To summarize, maximizing $I(\hat{y};\tilde{\mathbf{x}})$ is equivalent to minimizing the expected Kullback–Leibler (KL) divergence $\mathbb{E}\left(D_{\text{KL}}(\Phi_c(\tilde{\mathbf{x}});\Phi_c(\mathbf{x}))\right)$, and it provides an upper-bound to other loss functions such as the ℓ_1-loss $\mathbb{E}\left(\|\Phi_c(\tilde{\mathbf{x}}) - \Phi_c(\mathbf{x})\|_1^2\right)$.

To motivate the link between the compression term in the IB principle and the sparsity term $\mathbf{m}_1$ in the RDE objective, we begin by rewriting the mutual information $I(\mathbf{x};\tilde{\mathbf{x}})$ as:

$$I(\mathbf{x};\tilde{\mathbf{x}}) = H(\mathbf{x}) + H(\tilde{\mathbf{x}}) - H(\mathbf{x},\tilde{\mathbf{x}}) \tag{12}$$

Since $H(\mathbf{x})$ is constant when optimizing for a sparse mask $\mathbf{m} \in [0,1]^d$ within the IB principle, it can be dropped, yielding:

$$I(\mathbf{x};\tilde{\mathbf{x}}) = H(\tilde{\mathbf{x}}) - H(\mathbf{x},\tilde{\mathbf{x}}) \tag{13}$$

Using the conditional entropy definition, we get:

$$H(\mathbf{x},\tilde{\mathbf{x}}) = H(\tilde{\mathbf{x}}|\mathbf{x}) + H(\mathbf{x}) \tag{14}$$

Dropping $H(\mathbf{x})$ for the same reason as above, we obtain:

$$I(\mathbf{x};\tilde{\mathbf{x}}) = H(\tilde{\mathbf{x}}) - H(\tilde{\mathbf{x}}|\mathbf{x}) \tag{15}$$

Since $\tilde{\mathbf{x}} \subseteq \mathbf{x}$, the conditional entropy $H(\tilde{\mathbf{x}}|\mathbf{x})$ is small, leading to the approximation:

$$I(\mathbf{x};\tilde{\mathbf{x}}) \approx H(\tilde{\mathbf{x}}) \tag{16}$$

This demonstrates that the mutual information $I(\mathbf{x};\tilde{\mathbf{x}})$ can be effectively approximated by the entropy of the masked input, $H(\tilde{\mathbf{x}})$ (Strouse & Schwab, 2017; Kirsch et al., 2020). A practical way to minimize $H(\tilde{\mathbf{x}})$ is to reduce the number of explanatory variables, which corresponds to encouraging sparsity in $\mathbf{m}$ (Bang et al., 2021; Tao et al., 2020). This sparsity constraint is commonly represented by $\|\mathbf{m}\|_0$ and is approximated using $\|\mathbf{m}\|_1$ for optimization purposes. Thus, the sparsity term in the RDE objective conceptually aligns with the compression term in the IB principle

A.2 Expected RDE loss at initialization uniform versus gradient-based initialization

Let $\mathbf{x} \in \mathbb{R}^d$ denote the original input with dimensionality d. Further let $\mathbf{m}_{\text{unif}} \in [0,1]^d$ be the uniformly initialized mask where each m_i is independently sampled from $\mathcal{U}(0,1)$. Similarly, denote $\mathbf{m}_{\text{grad}}$ the gradient-based initialized mask as described in section 4, where we transformed the gradient values using a quantile transformation function so that $\mathbf{m}_{\text{grad}} \in [0,1]^d$. We aim to prove the following inequality:

$$\mathbb{E}_{\mathbf{m}_{\text{grad}},\mathbf{u}\sim V}\left[\mathcal{D}\left(\Phi_c(\tilde{\mathbf{x}}_{\text{grad}}), \Phi_c(\mathbf{x})\right) + \lambda\|\mathbf{m}_{\text{grad}}\|_1\right] \leq \mathbb{E}_{\mathbf{m}_{\text{unif}},\mathbf{u}\sim V}\left[\mathcal{D}\left(\Phi_c(\tilde{\mathbf{x}}_{\text{unif}}), \Phi_c(\mathbf{x})\right) + \lambda\|\mathbf{m}_{\text{unif}}\|_1\right] \tag{17}$$

Here, $\tilde{\mathbf{x}}_{\text{grad}} = \mathbf{m}_{\text{grad}} \odot \mathbf{x} + (1 - \mathbf{m}_{\text{grad}}) \odot \mathbf{u}$ and $\tilde{\mathbf{x}}_{\text{unif}} = \mathbf{m}_{\text{unif}} \odot \mathbf{x} + (1 - \mathbf{m}_{\text{unif}}) \odot \mathbf{u}$ are the distorted versions of the input $\mathbf{x} \in \mathbb{R}^d$ with random perturbation $\mathbf{u} \in \mathbb{R}^d$ drawn from a predefined probability distribution V such as Gaussian. Here, λ represents a hyperparameter, encouraging sparsity in $\mathbf{m}_{\text{grad}}$ and $\mathbf{m}_{\text{unif}}$.

Let further denote $\Phi_c(\mathbf{x})$ a classifier's prediction for the input $\mathbf{x} \in \mathbb{R}^d$ where $\Phi_c : \mathbb{R}^d \to [0,1]^c$ with $c \geq 1$ representing the number of classes. Assume that the classifier's prediction $\Phi_c(\mathbf{x})$ is differentiable and locally linear in the neighborhood of $\mathbf{x}$. Formally, this neighborhood is defined as the open ball $B_\epsilon(\boldsymbol{x}) = \{\tilde{\boldsymbol{x}} \in \mathbb{R}^d : \|\tilde{\boldsymbol{x}} - \boldsymbol{x}\| < \epsilon\}$, where $\epsilon > 0$ ensures that higher-order terms in the Taylor expansion of Φ_c are negligible.

Then, we can use the a first-order Taylor expansion around $\mathbf{x}$:

$$\Phi_c(\tilde{\mathbf{x}}) \approx \Phi_c(\mathbf{x}) + \nabla \mathbf{x}\Phi_c(\mathbf{x}) \cdot (\tilde{\mathbf{x}} - \mathbf{x}) \tag{18}$$

where $\nabla \mathbf{x}\Phi_c(\mathbf{x})$ represents the gradient of $\Phi_c(\mathbf{x})$ with respect to $\mathbf{x}$ and $\Delta\mathbf{x} = \tilde{\mathbf{x}} - \mathbf{x}$. The distortion $\mathcal{D}(\Phi_c(\tilde{\mathbf{x}}), \Phi_c(\mathbf{x}))$ can therefore be approximated, for a norm $\|\cdot\|_p$, as:

$$\mathcal{D}\left(\Phi_c(\tilde{\mathbf{x}}), \Phi_c(\mathbf{x})\right) = \|\Phi_c(\tilde{\mathbf{x}}) - \Phi_c(\mathbf{x}))\|_p \approx \|\nabla_{\mathbf{x}}\Phi_c(\mathbf{x}) \cdot \Delta\mathbf{x}\|_p \tag{19}$$

Using this, we get the following expression:

$$\mathbb{E}_{\mathbf{m}_{\text{grad}},\mathbf{u}\sim V}\left[\|\nabla_{\mathbf{x}}\Phi_c(\mathbf{x})\cdot\Delta\mathbf{x}\|_p+\lambda\|\mathbf{m}_{\text{grad}}\|_1\right]\leq\mathbb{E}_{\mathbf{m}_{\text{unif}},\mathbf{u}\sim V}\left[\|\nabla_{\mathbf{x}}\Phi_c(\mathbf{x})\cdot\Delta\mathbf{x}\|_p+\lambda\|\mathbf{m}_{\text{unif}}\|_1\right] \quad (20)$$

Since $\mathbf{m}_{\text{uniform}}, \mathbf{m}_{\text{grad}} \sim \mathcal{U}(0,1)^d$, the terms $\mathbb{E}_{\mathbf{m}_{\text{grad}}}\left[\lambda\|\mathbf{m}_{\text{grad}}\|_1\right]$ and $\mathbb{E}_{\mathbf{m}_{\text{unif}}}\left[\lambda\|\mathbf{m}_{\text{unif}}\|_1\right]$ simplify to $\frac{\lambda}{2}d$ as the expected value for each m_i is $\frac{1}{2}$. Both sparsity terms can be omitted from the comparison since they are equal. Therefore, we only need to show that:

$$\mathbb{E}_{\mathbf{m}_{\text{grad}},\mathbf{u}\sim V}\left[\|\nabla_{\mathbf{x}}\Phi_c(\mathbf{x})\cdot(\tilde{\mathbf{x}}_{\text{grad}}-\mathbf{x})\|_p\right]\leq\mathbb{E}_{\mathbf{m}_{\text{unif}},\mathbf{u}\sim V}\left[\|\nabla_{\mathbf{x}}\Phi_c(\mathbf{x})\cdot(\tilde{\mathbf{x}}_{\text{unif}}-\mathbf{x})\|_p\right]. \quad (21)$$

holds. By using the definitions for $\tilde{\mathbf{x}}_{\text{grad}}$ and $\tilde{\mathbf{x}}_{\text{unif}}$, we get the following after rearranging terms:

$$\tilde{\mathbf{x}}_{\text{grad}}-\mathbf{x}=(\mathbf{m}_{\text{grad}}-1)\odot(\mathbf{x}-\mathbf{u}), \quad (22)$$
$$\tilde{\mathbf{x}}_{\text{unif}}-\mathbf{x}=(\mathbf{m}_{\text{unif}}-1)\odot(\mathbf{x}-\mathbf{u}). \quad (23)$$

Hence, we can rearrange 21:

$$\mathbb{E}_{\mathbf{m}_{\text{grad}},\mathbf{u}\sim V}\left[\|\nabla_{\mathbf{x}}\Phi_c(\mathbf{x})\odot(\mathbf{m}_{\text{grad}}-1)\odot(\mathbf{x}-\mathbf{u})\|_p\right]\leq\mathbb{E}_{\mathbf{m}_{\text{unif}},\mathbf{u}\sim V}\left[\|\nabla_{\mathbf{x}}\Phi_c(\mathbf{x})\odot(\mathbf{m}_{\text{unif}}-1)\odot(\mathbf{x}-\mathbf{u})\|_p\right] \quad (24)$$

The key observation is that the gradient-based mask $\mathbf{m}_{\text{grad}}$ is constructed to minimize the distortion by assigning values of m_i close to 1 when $\nabla_{\mathbf{x}_i}\Phi_c(\mathbf{x})$ is large, thus reducing the effect of perturbations in important directions. On the other hand, the uniform mask $\mathbf{m}_{\text{unif}}$ is assigned independently of the gradient, making it less likely to reduce distortion effectively. Consequently, the gradient-based method is expected to achieve lower distortion compared to the uniform initialization *at initialization* by design, thereby validating the claim.

This completes the proof. □

A.3 Expected RDE loss at initialization all-ones versus gradient-based initialization

Let $\mathbf{m}_{\text{ones}}$ denote the mask initialized with all ones, i.e. $\mathbf{m}_{\text{ones}} = \mathbf{1}_d$ where $\mathbf{1}_d$ is a vector of ones with dimension d. Let further denote $\Phi_c(\mathbf{x})$ a classifier's prediction for the input $\mathbf{x}\in\mathbb{R}^d$ where $\Phi_c:\mathbb{R}^d\to[0,1]^c$ with $c\geq 1$ representing the number of classes. Given the function $\mathcal{D}:[0,1]^c\times[0,1]^c\to\mathbb{R}_+$ that measures the distortion between the classifier's prediction for the original input $\mathbf{x}$ and the 'distorted' input $\tilde{\mathbf{x}}$, the all-ones initialization results in $\mathcal{D}\left(\Phi_c(\mathbf{x}),\Phi_c(\tilde{\mathbf{x}})\right)=0$ as we have $\mathbf{x}=\tilde{\mathbf{x}}$. Furthermore, let $\mathbf{m}_{\text{grad}}\sim\mathcal{U}(0,1)^d$ denote the mask initialized using the gradient-based scheme as described in section 4. We aim to show that

$$\mathbb{E}_{\mathbf{m}_{\text{grad}},\mathbf{u}\sim V}\left[\mathcal{D}\left(\Phi_c(\tilde{\mathbf{x}}),\Phi_c(\mathbf{x})\right)+\lambda\|\mathbf{m}_{\text{grad}}\|_1\right]\leq\lambda\|\mathbf{m}_{\text{ones}}\|_1 \quad (25)$$

holds, where λ represents a hyperparameter balancing the trade-off between distortion and sparsity. Since $\mathbf{m}_{\text{ones}}$ is initialized with all ones, we have $\|\mathbf{m}_{\text{ones}}\|_1=d$. For $\mathbf{m}_{\text{grad}}\sim\mathcal{U}(0,1)^d$, the expected value of the l_1-norm is given by $\mathbb{E}_{\mathbf{m}_{\text{grad}}}\left[\|\mathbf{m}_{\text{grad}}\|_1\right]=\frac{d}{2}$. Therefore, we can simplify the inequality as follows:

$$\mathbb{E}_{\mathbf{m}_{\text{grad}},\mathbf{u}\sim V}\left[\mathcal{D}\left(\Phi_c(\tilde{\mathbf{x}}),\Phi_c(\mathbf{x})\right)\right]+\frac{\lambda}{2}d\leq\lambda d \quad (26)$$
$$\mathbb{E}_{\mathbf{m}_{\text{grad}},\mathbf{u}\sim V}\left[\mathcal{D}\left(\Phi_c(\tilde{\mathbf{x}}),\Phi_c(\mathbf{x})\right)\right]\leq\frac{\lambda}{2}d \quad (27)$$

If we use a norm $\|\cdot\|_p$ with $p\geq 1$ as a choice for the distortion function $\mathcal{D}$, we can upper bound the expected distortion on the left-hand side of equation in the following way:

$$\mathbb{E}_{\mathbf{m}_{\text{grad}},\mathbf{u}\sim V}\left[\mathcal{D}\left(\Phi_c(\tilde{\mathbf{x}}),\Phi_c(\mathbf{x})\right)\right] \leq \left(1+(2-1)\mathbb{1}_{c\geq 2}\right)^{\frac{1}{p}} \leq \frac{\lambda}{2}d \tag{28}$$

as $\Phi_c : \mathbb{R}^d \to [0,1]^c$ for the classification setting we consider, where $\sum_{i=1}^{c}\Phi_{c,i}(x) = 1$ for any input $x \in \mathbb{R}^d$. Therefore, the gradient-based mask initialization has a lower expected loss in terms of distortion and sparsity *at initialization*, if

$$\frac{2\left(1+(2-1)\mathbb{1}_{c\geq 2}\right)^{\frac{1}{p}}}{d} \leq \lambda \tag{29}$$

holds. In high-dimensional settings, which are typical for deep learning models and, consequently, for mask-based explanation methods, this condition is generally satisfied.

This completes the proof. □

B Pseudocode

B.1 General version of the StartGrad Algorithm

While we assumed in the pseudocode for Algorithm 1 that we operate in the same input space $\mathcal{X}$, we outline in Algorithm 2 a more general version of the StartGrad algorithm, where we first use an invertible mapping $\mathcal{F}$.

Algorithm 2: General version: Gradient-based Mask Initialization (StartGrad)

```
Input          : Pre-trained classifier $\Phi_c$, input samples $\{\boldsymbol{x}_i\}_{i=1}^{N}$ with $\boldsymbol{x}_i \in \mathbb{R}^d$, quantile
                 transformation $\mathcal{T} : \mathbb{R}_+^d \to [0,1]^d$, invertible and differentiable function
                 $\mathcal{F} : \mathbb{R}^d \to \mathbb{R}^k$
Hyperparameters: Output distribution $\mathcal{U}(0,1)$, number of quantiles (bins) for $\mathcal{T}$
Output         : Masks $\mathbf{M} \in [0,1]^{N \times d}$
Initialize mask list $\mathbf{M} := []$ and quantile transformer $\mathcal{T}$;
for $i \leftarrow 1$ to $N$ do
    $c_i \leftarrow \arg\max(\Phi_c(\boldsymbol{x}_i))$ // Class prediction
    $\mathbf{S}_i \leftarrow \left|\nabla_{\mathcal{F}(\boldsymbol{x}_i)} \Phi_c(\boldsymbol{x}_i)\right|_c$ // Gradient magnitudes
    $\mathbf{m}_i \leftarrow \mathcal{T}(\mathbf{S}_i)$ // Quantile transform
    Append $\mathbf{m}_i$ to $\mathbf{M}$;
end for
return $\mathbf{M}$
```

Figure 2 visually illustrates the pseudocode.

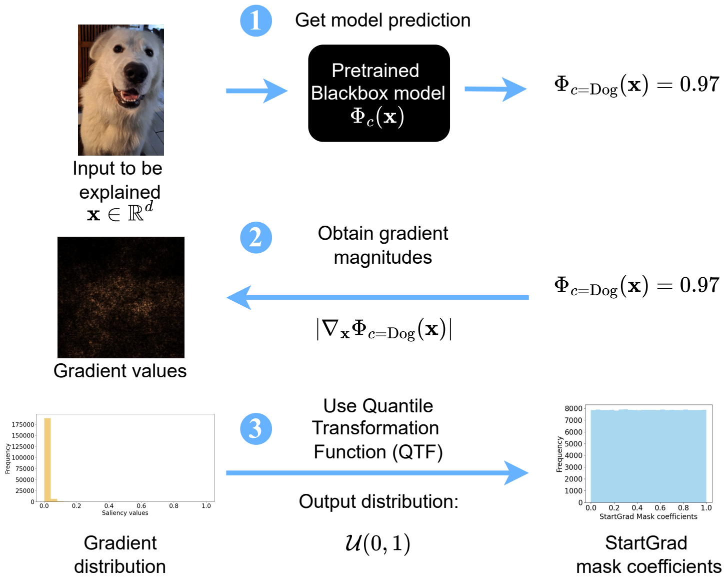

1
Get model prediction
Pretrained Blackbox model $\Phi_c(\mathbf{x})$
$\Phi_{c=\text{Dog}}(\mathbf{x}) = 0.97$
Input to be explained $\mathbf{x} \in \mathbb{R}^d$
2
Obtain gradient magnitudes
$\Phi_{c=\text{Dog}}(\mathbf{x}) = 0.97$
$|\nabla_{\mathbf{x}} \Phi_{c=\text{Dog}}(\mathbf{x})|$
Gradient values
3
Use Quantile Transformation Function (QTF)
Output distribution:
$\mathcal{U}(0,1)$
Frequency
175000
150000
125000
100000
75000
50000
25000
0
0.0
0.2
0.4
0.6
0.8
1.0
Saliency values
Gradient distribution
Frequency
8000
7000
6000
5000
4000
3000
2000
1000
0
0.0
0.2
0.4
0.6
0.8
1.0
StartGrad Mask coefficients
StartGrad mask coefficients

Figure 2: Visual illustration of the StartGrad algorithm.

C Experimental Settings and Details

C.1 Vision

Here we describe the explanation methods used for the vision part of the experiments in more detail. We use the same notation as already introduced in section 3.

For all vision experiments, we optimize the mask coefficients for 300 steps, using an Adam optimizer (Kingma & Ba, 2014) with learning rate of 10^{-1}.

C.1.1 PixelMask

The objective for the PixelMask model as first introduced by Fong & Vedaldi (2017) is:

$$\min_{\mathbf{m}\in[0,1]^k} E_{\mathbf{u}\sim V}\left[\mathcal{D}\left(\Phi_c(\tilde{\mathbf{x}}), \Phi_c(\mathbf{x})\right)\right] + \lambda_1\|\mathbf{m}\|_1 \tag{30}$$

We set $\lambda_1 = 1$, and estimate the expectation via Monte Carlo sampling across 16 samples, where we draw the perturbation u from an uniform distribution. In particular, we use an uniform noise from $[\mu - \sigma, \mu + \sigma]$, where μ and σ are the empirical mean and standard deviation of the pixel values of the image as it is been done in (Kolek et al., 2022; 2023).

For the PixelMask (Fong & Vedaldi, 2017) model, we use 65,536 bins for the QTF function in Algorithm 1 which equals the width and the height (256 by 256) of the images used for the pretrained vision models.

C.1.2 ShearletX & WaveletX

For the ShearletX (Kolek et al., 2023) and WaveletX model (Kolek et al., 2023), we utilized the original implementations and follow their experimental setting closely.

With the same notation as used in section 3, Kolek et al. (2023) extend equation 30 in the following manner:

$$\min_{\mathbf{m}\in[0,1]^k} E_{\mathbf{u}\sim V}\left[\mathcal{D}\left(\Phi_c(\tilde{\mathbf{x}}), \Phi_c(\mathbf{x})\right)\right] + \lambda_1\|\mathbf{m}\|_1 + \lambda_2\|\mathcal{F}^{-1}\left(\mathbf{m}\odot\mathcal{F}(\mathbf{x})\right)\|_1 \tag{31}$$

where $\tilde{\mathbf{x}} := \mathcal{F}^{-1}(\mathbf{m}\odot\mathcal{F}(\mathbf{x})+(1-\mathbf{m})\odot\mathbf{u})$ denotes now the distorted input image after applying the mask in the transformed space. The WaveletX model uses for the $\mathcal{F}$ the discrete wavelet transform (DWT), whereas the ShearletX makes use of the digital shearlet transform (Kutyniok et al., 2016; Kolek et al., 2023). For the WaveletX we use the daubechies-3 as a mother wavelet with five scales.

The expectation in 31 is approximated using Monte Carlo sampling using 16 samples where the perturbation for scale a and shearing s is sampled uniformly from $[\mu_{a,s} - \sigma_{a,s}, \mu_{a,s} + \sigma_{a,s}]$. Here $\sigma_{a,s}$ and $\mu_{a,s}$ are the empirical standard deviation and mean of the image's shearlet coefficients at scale a and shearing s. In line with Kolek et al. (2023), we set for ShearletX $\lambda_1 = 1$ and $\lambda_2 = 2$. For WaveletX we set $\lambda_1 = 1$ and $\lambda_2 = 10$. We use the mean-squared error as a choice for the distortion function $\mathcal{D}$.

For the ShearletX and WaveletX (Kolek et al., 2023), we use 10,000 as the number of bins for the former, and the following bins for each of the scales, 144, 16900, 4489, 1296, 400, 144 for the latter.

C.2 Vision evaluation metrics

C.2.1 CP-L1 and CP-Pixel score

As outlined in section 5, we use the conciseness-preciseness (CP) Pixel and L1 score first introduced by Kolek et al. (2023) as quantitative performance metrics which are defined as follows:

$$\text{CP} = \frac{\text{Retained class probability}}{\text{Retained image information}} \tag{32}$$

where the numerator is defined by:

$$\text{Retained class probability} = \frac{\Phi_c(\tilde{\mathbf{x}})}{\Phi_c(\mathbf{x})} \tag{33}$$

while the denominator in case of the CP-Pixel is defined as:

$$\text{Retained image information Pixel} = \frac{\|\mathbf{m} \odot \mathbf{x}\|_1}{\|\mathbf{x}\|_1} \tag{34}$$

whereas for the CP-L1 we have:

$$\text{Retained image information L1} = \frac{\|\mathbf{m} \odot \mathbf{c}\|_1}{\|\mathbf{c}\|_1} \tag{35}$$

where $\mathbf{c} \in \mathbb{R}^k$ represents the coefficients of the latent space such as the shearlet or wavelet space. Note that for the PixelMask model, the CP-L1 and CP-Pixel metrics are identical as the masking is only applied in the pixel space.

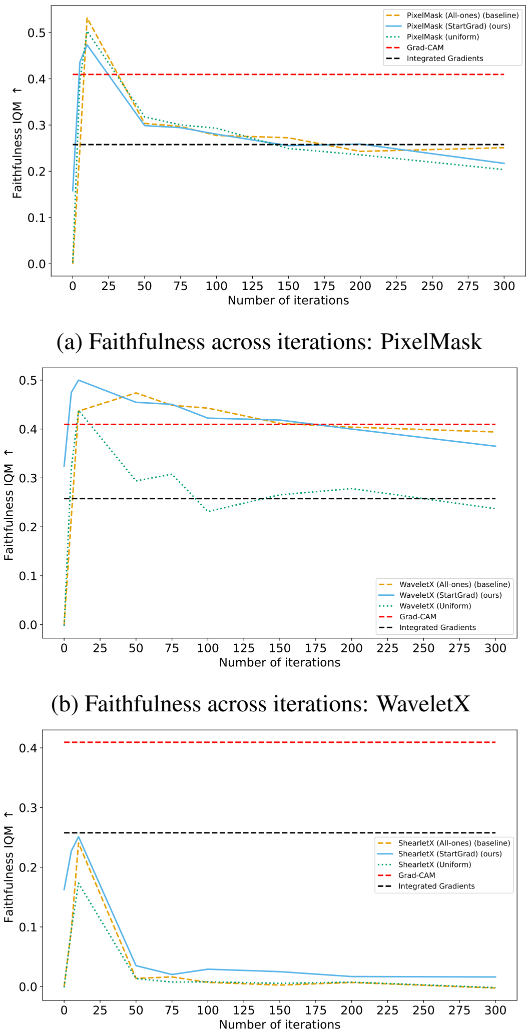
Given the aforementioned formulations of the CP scores, one can notice that these scores strive to balance both, preciseness and sparsity of the explanation. Higher scores indicate superior masks (Kolek et al., 2023).

C.2.2 Justification of CP-L1 and CP-Pixel score as opposed to deletion and insertion score

Attribution methods like Integrated Gradients (Sundararajan et al., 2017) and SmoothGrad (Smilkov et al., 2017) are typically evaluated using insertion and deletion scores. Unlike the mask-based explanation methods in this study, they provide a clear feature importance ranking, usually in pixel space. In contrast, mask-based methods do not inherently rank coefficient relevance but instead optimize a binary, sparse mask using a relaxed l_1 norm to approximate the ideal l_0 norm. This binary nature complicates insertion and deletion scores, which require an ordered ranking. This issue was previously highlighted by Kolek et al. (2023), who introduced CP-scores; we refer to their work for further discussion.

To highlight the issues with using the insertion and deletion metrics to evaluate mask-based explanation methods, we computed the faithfulness scores which is the difference between the former two for PixelMask (Fong & Vedaldi, 2017), WaveletX, and ShearletX (Kolek et al., 2023) across multiple iterations, as shown in the following figure. As we can see from the figure 3 the faithfulness scores decreases with increasing iteration steps across all three methods and all initializations and reach their maximum scores at very early iterations steps which is in sharp contrast to the results obtained using the CP-scores and also counterintuitive as this implies that the learned mask gets less faithful over time. This illustrates the problem of faithfulness score which presumes an inherent ordering of the mask values, even though the mask-based explanation methods optimize a relaxed binary problem with no inherent ordering. For comparison reasons, we included two non-masked based attribution methods Integrated Gradients (Sundararajan et al., 2017) and GradCAM (Selvaraju et al., 2017)

For completeness, Table 4 includes a comparison between the two aforementioned gradient-based methods and the mask-based explanation methods used in this study using the CP-Pixel and CP-L1 evaluation metric. The results clearly demonstrate that mask-based explanation methods consistently outperform non-mask-based methods across all iteration steps. However, it is important to note that mask-based methods, with their learned binary mask, are explicitly designed to balance distortion and sparsity, giving them an advantage in this trade-off.



(a) Faithfulness across iterations: PixelMask

(b) Faithfulness across iterations: WaveletX

(c) Faithfulness across iterations: ShearletX

Figure 3: Faithfulness scores across iteration for the PixelMask (Fong & Vedaldi, 2017) (a), WaveletX (Kolek et al., 2023) (b), and ShearletX (Kolek et al., 2023) (c) models, comparing three initialization methods: StartGrad (ours) (solid line), All-ones (dashed line), and Uniform (dotted line). For comparison reasons, we added two non-masked based explanation methods, i.e. Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017) and Integrated Gradients (Sundararajan et al., 2017). As these methods are not iteratively optimized, the faithfulness values for these methods remain constant. The curves represent the interquartile mean (IQM) across 100 randomly selected ImageNet validation samples (Deng et al., 2009), evaluated with a pretrained ResNet18 classifier (He et al., 2016). ↑ indicates that higher is better. We can see that the faithfulness score decreases across all methods and initializations over the course of optimization which is counterintuitive and in contrast to the CP-scores, thereby illustrating the problem associated with using faithfulness scores as an evaluation metric for mask-based explanation methods that do not inherently order their mask-values according to importance.

Table 4: Interquartile mean score (IQM) for the PixelMask (Fong & Vedaldi, 2017), WaveletX (Kolek et al., 2023), ShearletX (Kolek et al., 2023), Integrated Gradients (IG) (Sundararajan et al., 2017), Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017) of the CP-Pixel and CP-L1 score across iteration steps. All mask-based methods are initialized with StartGrad. To map the gradient-based attributions (IG, GradCAM) into the desired range of $[0, 1]$, either a min-max normalization (Min-Max) or a quantile transformation function (QTF) were used. All experiments use 500 randomly selected ImageNet (Deng et al., 2009) samples evaluated on a pretrained ResNet18 (He et al., 2016) classifier. For each metric, ↑ indicates that higher is better, and ↓ that lower is better. Values that are two standard errors away from the second-best values are highlighted in bold. As IG (Sundararajan et al., 2017) and GradCAM (Selvaraju et al., 2017) are both methods that are not optimized iteratively, their scores remain constant across the shown iteration steps. The CP-Pixel and CP-L1 scores differ only for the WaveletX (Kolek et al., 2023) and ShearletX (Kolek et al., 2023), as only these methods operate also in a latent space.

		Iteration steps		
Metric	Initialization	50	100	300
CP-Pixel ↑	PixelMask	37.395 ± 1.119	51.217 ± 1.478	65.766 ± 1.767
	WaveletX	48.973 ± 0.604	64.551 ± 0.809	77.534 ± 0.976
	ShearletX	$\mathbf{216.908 \pm 6.997}$	$\mathbf{359.803 \pm 9.675}$	$\mathbf{481.314 \pm 11.829}$
	IG (Min-Max)	3.237 ± 0.015	3.237 ± 0.015	3.237 ± 0.015
	IG (QTF)	2.037 ± 0.001	2.037 ± 0.001	2.037 ± 0.001
	GradCAM (Min-Max)	2.841 ± 0.019	2.841 ± 0.019	2.841 ± 0.019
	GradCAM (QTF)	1.987 ± 0.004	1.987 ± 0.004	1.987 ± 0.004
CP-L1 ↑	PixelMask	37.395 ± 1.119	51.217 ± 1.477	65.766 ± 1.767
	WaveletX	11.208 ± 0.128	14.505 ± 0.173	17.484 ± 0.204
	ShearletX	$\mathbf{123.791 \pm 4.760}$	$\mathbf{199.193 \pm 6.384}$	$\mathbf{270.438 \pm 7.931}$
	IG (Min-Max)	3.237 ± 0.015	3.237 ± 0.015	3.237 ± 0.015
	IG (QTF)	2.037 ± 0.001	2.037 ± 0.001	2.037 ± 0.001
	GradCAM (Min-Max)	2.841 ± 0.019	2.841 ± 0.019	2.841 ± 0.019
	GradCAM (QTF)	1.987 ± 0.004	1.987 ± 0.004	1.987 ± 0.004

C.3 Time Series

C.3.1 State Dataset

The dataset for the synthetic state dataset as introduced by Tonekaboni et al. (2020) generates data according to a 2-state hidden Markov model (HMM). Let $t \in [t : T]$ represent the discrete time steps and $s_t \in \{0, 1\}$ denote the specific states. For this dataset, the input $\mathbf{x} \in \mathbb{R}^3$ is generated using a multivariate Normal distribution $x_t \sim \mathcal{N}(\mu_{s_t}, \Sigma_{s_t})$, where the mean μ_{s_t} and the variance-covariance matrix Σ_{s_t} are dependent on the state at s_t. In particular, we set $\mu_0 = [0.1, 1.6, 0.5]$ and $\mu_1 = [-0.1, -0.4, -1.5]$. The variance-covariance matrix is given by:

$$\Sigma = \begin{bmatrix} 0.8 & 0.01 & 0.01 \\ 0.01 & 0.8 & 0.0 \\ 0.01 & 0.0 & 0.8 \end{bmatrix}$$

The first feature dimension is irrelevant for the label y_t, whereas the label is distributed as:

$$y_t \sim \text{Bernoulli}(p_t)$$

$$p_t = \begin{cases} (1 + \exp[-2x_{2,t}])^{-1} & \text{if } s_t = 0 \\ (1 + \exp[-2x_{3,t}])^{-1} & \text{if } s_t = 1 \end{cases} \tag{36}$$

The transition probabilities are given by:

$$T = \begin{bmatrix} 0.1 & 0.9 \\ 0.01 & 0.9 \end{bmatrix}$$

with the starting state at $t = 0$ being determined with equal probability, i.e. $\pi = [\frac{1}{2}, \frac{1}{2}]$.

For the experiments, we generate 1,000 time series, 800 for training, and 200 for testing.

C.3.2 Switch-Feature Dataset

Following the setup introduced by Tonekaboni et al. (2020), the switch-feature dataset uses a Gaussian process mixture in place of the multivariate normal distribution with means μ

$$\mu = \begin{bmatrix} 0.8 & 0.5 & 0.2 \\ 0.0 & 1.0 & 0.0 \\ 0.2 & 0.2 & 0.8 \end{bmatrix}$$

The Gaussian process over time is governed by a RBF kernel with $\gamma = 0.2$, with marginal feature set to 0.1 for all states. In a similar vein to the state dataset, the label is distributed as:

$$y_t \sim \text{Bernoulli}(p_t)$$

$$p_t = \begin{cases} (1 + \exp[-3x_{1,t}])^{-1} & \text{if } s_t = 0 \\ (1 + \exp[-3x_{2,t}])^{-1} & \text{if } s_t = 1 \\ (1 + \exp[-3x_{3,t}])^{-1} & \text{if } s_t = 2 \end{cases} \tag{37}$$

The transition probabilities are given by:

$$T = \begin{bmatrix} 0.95 & 0.02 & 0.3 \\ 0.02 & 0.95 & 0.03 \\ 0.03 & 0.02 & 0.95 \end{bmatrix}$$

with the starting state at $t = 0$ being determined with equal probability, i.e. $\pi = [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$.

For the experiments, we generate 1,000 time series, 800 for training, and 200 for testing.

C.4 ExtremalMask

ExtremalMask (Enguehard, 2023) is a mask-based explanation method specifically designed for multivariate time-series data and is defined by:

$$\min_{\mathbf{m} \in [0,1]^{dxT}, \theta} \mathcal{D}\left(\Phi_c(\tilde{\mathbf{x}}), \Phi_c(\mathbf{x})\right) + \lambda_1 \|\mathbf{m}\|_1 + \lambda_2 \|f_\theta(\mathbf{x})\|_1 \tag{38}$$

where $\tilde{\mathbf{x}} = \mathbf{m} \odot \mathbf{x} + (1 - \mathbf{m}) \odot f_\theta(\mathbf{x})$ and $\mathbf{x} \in \mathbb{R}^{dxT}$ where d is the input dimensionality, and T represents the time dimension, respectively.

Equation 38 denotes the preservation game which aims to find the sparsest mask that preserves the most class probability. In the deletion game formulation of ExtremalMask (Enguehard, 2023), the objective function is defined as:

$$\min_{\mathbf{m} \in [0,1]^{dxT}, \theta} \mathcal{D}\left(\Phi_c(\tilde{\mathbf{x}}), \Phi_c(\mathbf{0})\right) + \lambda_1 \|1 - \mathbf{m}\|_1 + \lambda_2 \|f_\theta(\mathbf{x})\|_1 \tag{39}$$

In line with the experiments conducted in Enguehard (2023), we use a one-layer, bidirectional GRU (Cho et al., 2014) model with hidden dimensionality of 200 as choice for the f_θ. Furthermore, we use the Adam (Kingma & Ba, 2014) optimizer with learning rate 10^{-1}. We run the ExtremalMask (Enguehard, 2023) model for 500 iterations. We use the mean-squared error as a choice for the distortion function.

As opposed to the original formulation of the distortion objective 39, we use a slightly modified version of the objective which has shown to significantly boost the performance (Bant et al., 2024):

$$\min_{\mathbf{m}\in[0,1]^{dxT},\theta} \mathcal{D}\left(\Phi_c(\tilde{\mathbf{x}}), \Phi_c(\mathbf{0})\right) + \lambda_1 \|1 - \mathbf{m}\|_1 + \lambda_2 \|f_\theta(\mathbf{x}) - \mathbf{x}\|_2^2 \tag{40}$$

For all the experiments, we fit a single layer one-directional GRU (Cho et al., 2014) as a baseline classification model with hidden dimension 200 and train it for 50 epochs, batch size 128, learning rate of $1e-4$ with the Adam (Kingma & Ba, 2014) optimizer. We minimize the cross-entropy loss for training the GRU model.

For both time-series datasets, we use 600 (which is equal to the signal length times the feature dimension) as the number of bins for the Quantile Transformation Function (QTF) in our StartGrad algorithm 2.

C.5 Time-series evaluation metrics

Given that the ground truth salient features are know in our synthetic dataset, we use the two standard performance metric area under recall (AUR) and area under precision (AUP). We further evaluate StartGrad using two additional metrics, information and mask entropy, with

$$\mathrm{I}_M(A) = -\sum_{t,i\in A} \log(1 - m_{t,i}) \tag{41}$$

for $A \subseteq [1:T] \times [1:d_X]$ defining the information, and

$$\mathrm{S}_M(A) = -\sum_{t,i\in A} m_{t,i}\log(m_{t,i}) + (1 - m_{t,i})\log(1 - m_{t,i}) \tag{42}$$

the entropy respectively (Crabbé & Van Der Schaar, 2021). A higher score indicates a better score for information, whereas a lower score indicates a better performance with respect to entropy.

D Vision Experiments

D.1 Distribution of absolute gradient values

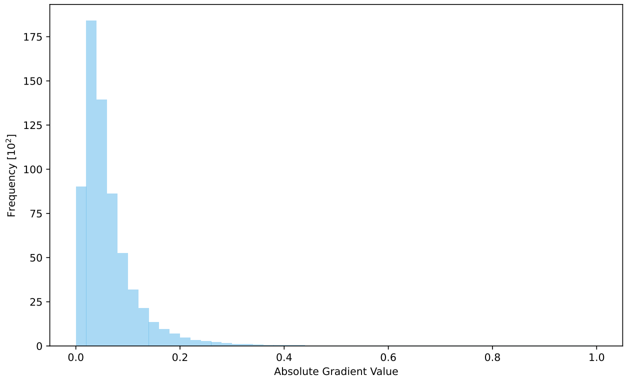


Figure 4: Absolute gradient values for the input pixels for 100 randomly selected validation ImageNet (Deng et al., 2009) samples using a pretrained ResNet18 (He et al., 2016) model. Here, the gradient was taken with respect to the class prediction.

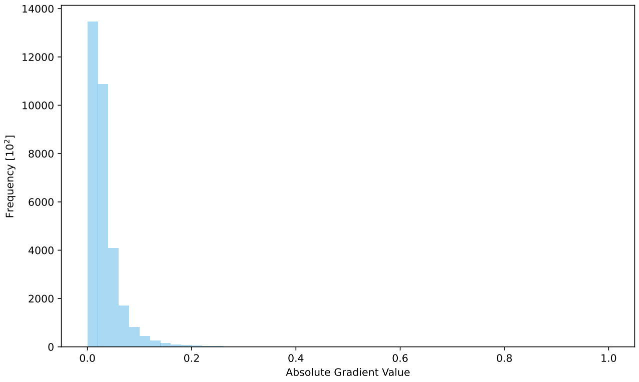


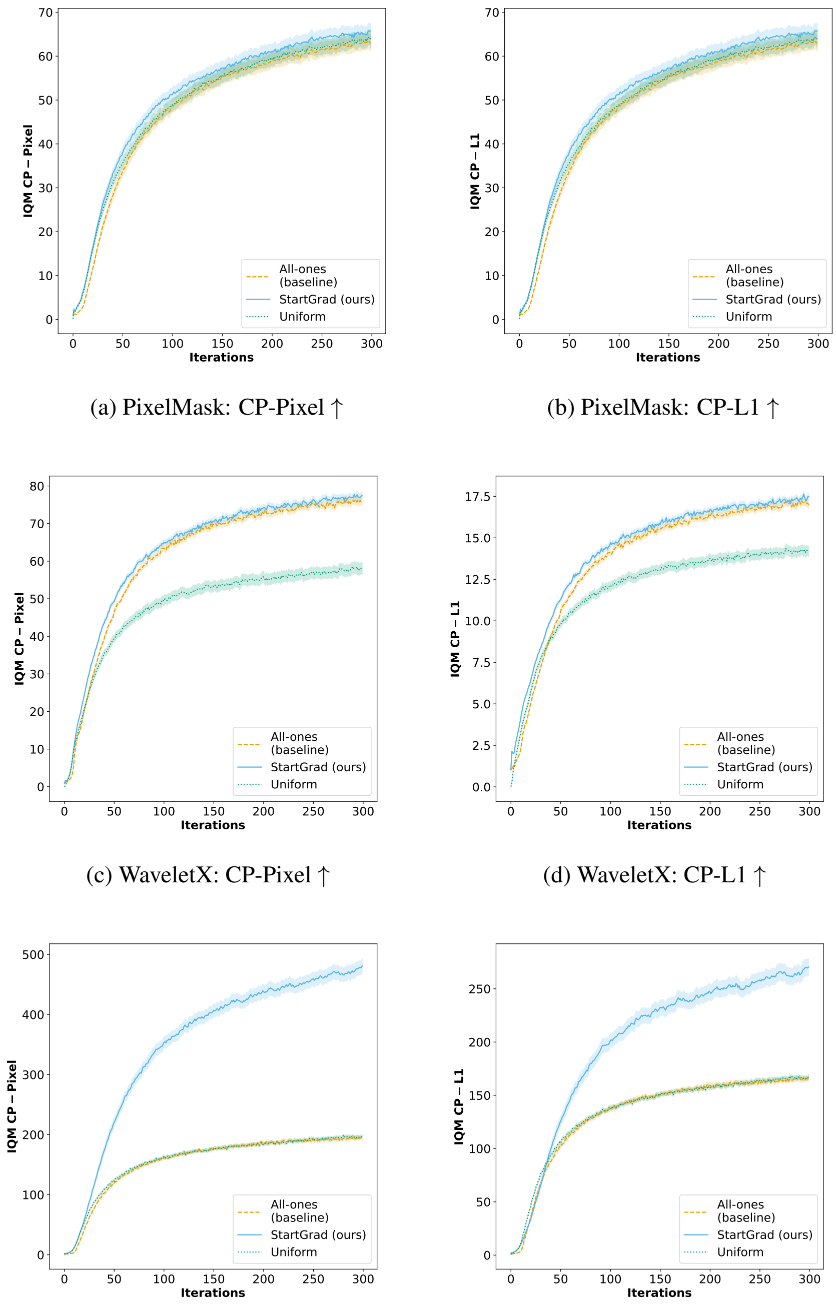
Figure 5: Absolute gradient values in the shearlet domain for the input pixels for 100 randomly selected validation ImageNet (Deng et al., 2009) samples using a pretrained ResNet18 (He et al., 2016) model. Here, the gradient was taken with respect to the class prediction.

D.2 Runtime analysis

On a Apple Macbook Pro with a M1 chip, we ran the StartGrad algorithm for all three vision models, i.e. PixelMask (Fong & Vedaldi, 2017), WaveletX (Kolek et al., 2023) and ShearletX (Kolek et al., 2023) across 100 randomly selected ImageNet validation samples (Deng et al., 2009) and obtain the following average runtime (in seconds):

- PixelMask: 3.62s
- WaveletX: 0.72s
- ShearletX: 1.31s

D.3 Additional Figures for Vision Experiments



(a) PixelMask: CP-Pixel ↑

(b) PixelMask: CP-L1 ↑

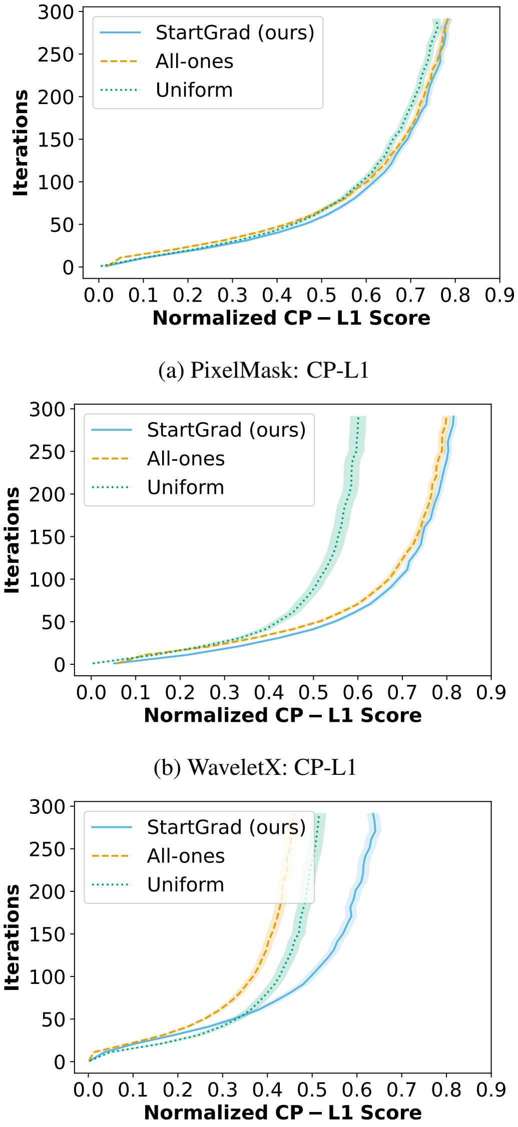
(c) WaveletX: CP-Pixel ↑

(d) WaveletX: CP-L1 ↑

(e) ShearletX: CP-Pixel ↑

(f) ShearletX: CP-L1 ↑

Figure 6: Comparison of StartGrad initialization (ours) with standard baseline initialization schemes for the PixelMask (Fong & Vedaldi, 2017) (first row), WaveletX (Kolek et al., 2023) (second row) and ShearletX (Kolek et al., 2023) (third row). Baseline refers to the originally used initialization scheme for the respective mask explanation method. The solid line represents the average of the interquartile mean (IQM) performance across 500 randomly selected validation ImageNet (Deng et al., 2009) samples, while the shaded area denotes the standard errors respectively. ↑ indicates that higher is better, and ↓ that lower is better. We use a pretrained ResNet18 (He et al., 2016) classifier.



(a) PixelMask: CP-L1

(b) WaveletX: CP-L1

(c) ShearletX: CP-L1

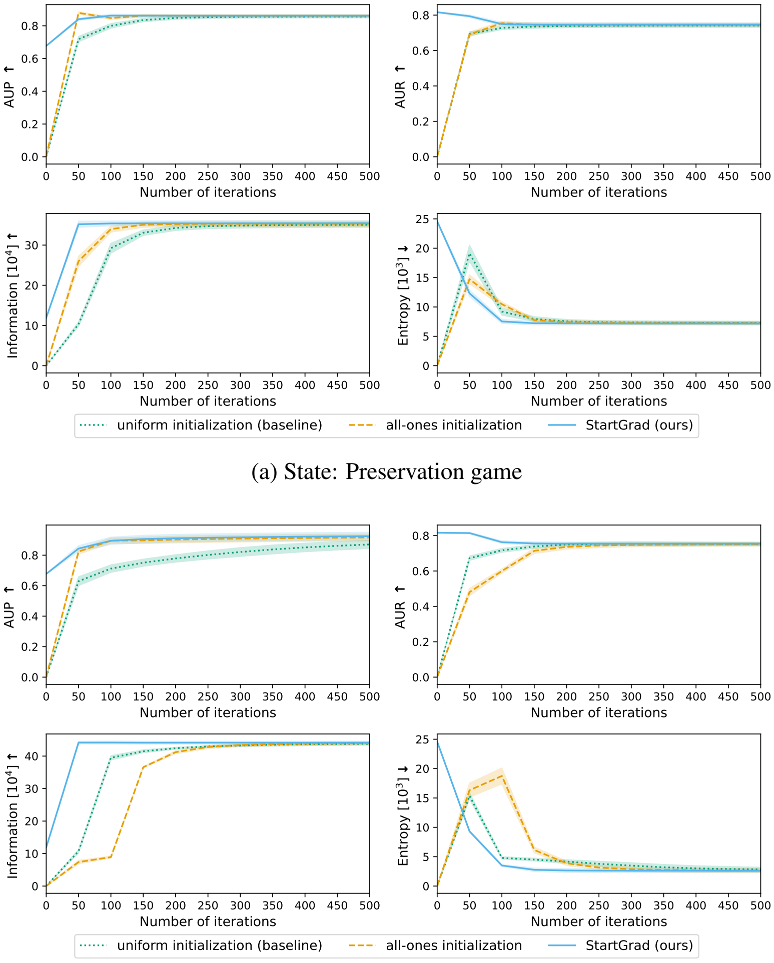
Figure 7: Normalized CP-L1 score vs. iteration steps for the PixelMask (Fong & Vedaldi, 2017) (a), WaveletX (Kolek et al., 2023) (b), and ShearletX (Kolek et al., 2023) (c) models, comparing three initialization methods: StartGrad (ours) (solid line), All-ones (dashed line), and Uniform (dotted line). The curves represent the average number of iteration steps needed to reach a normalized target CP-L1 score. This target score is defined as the highest CP-L1 value achieved across all three initialization methods for each model, then normalized by the overall maximum score observed. The shaded regions represent the standard error across 500 randomly selected ImageNet validation samples (Deng et al., 2009)xs, with a pretrained ResNet18 (He et al., 2016) classifier. StartGrad enables all three models to achieve target scores with fewer iterations compared to standard initialization schemes, with the effect being most notable for ShearletX (Kolek et al., 2023).

E Time-Series Experiments

In this section, the main results from the time-series experiments are presented which were omitted from the main text for the sake of brevity. A complete description of the datasets, models, hyperparameters and metrics used can be found in the Appendix C.3.

E.1 Figures

E.1.1 State Dataset

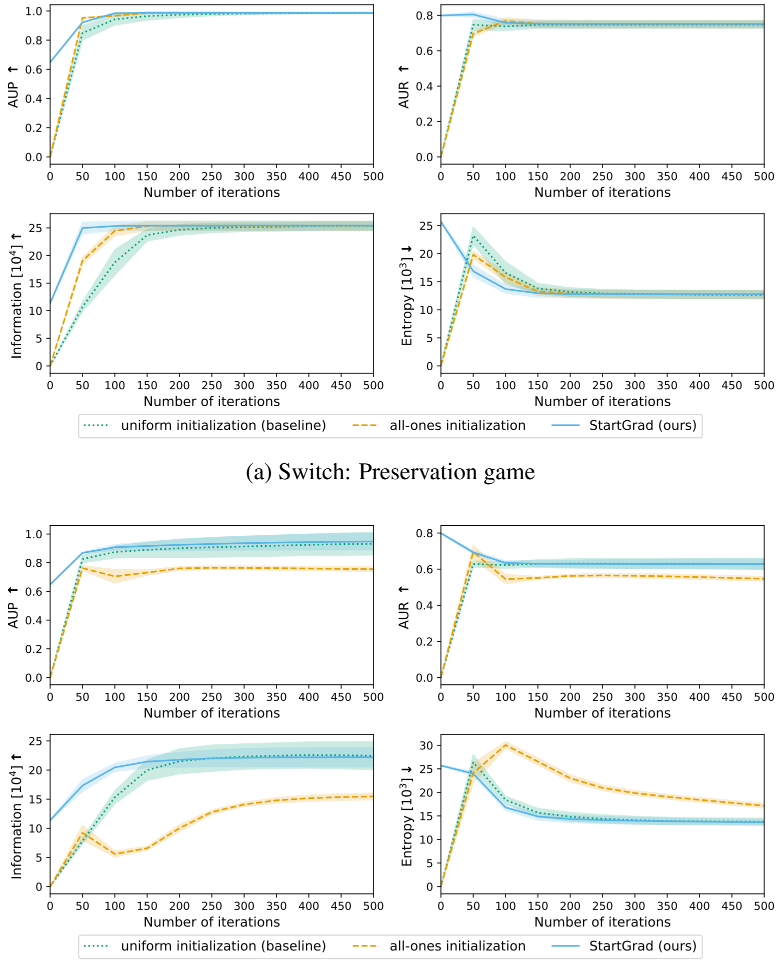


(a) State: Preservation game

(b) State: Deletion game

Figure 8: Comparison of the performance of the StartGrad (ours), all-ones and uniform (baseline) initialization on the state synthetic dataset. A one-layer GRU (Cho et al., 2014) is employed as a classifier, whereas the ExtremalMask (Enguehard, 2023) is used as a mask-based explanation method. The first panel (a) shows the results for the preservation game formulation of the ExtremalMask (Enguehard, 2023) model, where the goal is to preserve class probability as possible, while returning a sparse mask. The second panel (b) shows the results for the deletion game formulation of the ExtremalMask (Enguehard, 2023) model, where the objective is to find the least possible changes to the original input that distorts the class probability maximally. The solid line represents the average across five runs, with the shaded area highlighting the standard deviation. The preservation and deletion games objective versions are shown for both datasets. For each metric, ↑ indicates that higher is better, and ↓ that lower is better.

E.1.2 Switch-feature dataset



(a) Switch: Preservation game

(b) Switch: Deletion game

Figure 9: Comparison of the performance of the StartGrad (ours), all-ones and uniform (baseline) initialization on the switch-feature synthetic dataset. A one-layer GRU (Cho et al., 2014) is employed as a classifier, whereas the ExtremalMask (Enguehard, 2023) is used as a mask-based explanation method. The first panel (a) shows the results for the preservation game formulation of the ExtremalMask (Enguehard, 2023) model, where the goal is to preserve class probability as possible, while returning a sparse mask. The second panel (b) shows the results for the deletion game formulation of the ExtremalMask (Enguehard, 2023) model, where the objective is to find the least possible changes to the original input that distorts the class probability maximally. The solid line represents the average across five runs, with the shaded area highlighting the standard deviation. The preservation and deletion games objective versions are shown for both datasets. For each metric, ↑ indicates that higher is better, and ↓ that lower is better.

E.2 Average performance results

E.2.1 State Dataset

Table 5: Average performance for the ExtremalMask (Enguehard, 2023) explanation method (preservation game objective) across iteration steps for the state dataset when initialized with StartGrad, all-ones and uniformly. The reported numbers are the mean and standard deviation across five folds. For each metric, ↑ indicates that higher is better, and ↓ that lower is better. Baseline refers to the initialization scheme originally used for the respective mask explanation method. Outcomes that are one standard deviation away from the second-best one are highlighted in bold.

		Iteration steps			
Metric	Initialization	50	100	300	500
AUP ↑	Uniform (baseline)	0.719 ± 0.016	0.800 ± 0.013	0.855 ± 0.008	0.856 ± 0.008
	All-ones	**0.879 ± 0.006**	0.847 ± 0.008	0.860 ± 0.007	0.860 ± 0.007
	StartGrad (ours)	0.840 ± 0.007	**0.862 ± 0.007**	0.861 ± 0.008	0.861 ± 0.007
AUR ↑	Uniform (baseline)	0.694 ± 0.011	0.728 ± 0.009	0.741 ± 0.010	0.742 ± 0.010
	All-ones	0.691 ± 0.013	**0.755 ± 0.010**	0.745 ± 0.010	0.745 ± 0.010
	StartGrad (ours)	**0.794 ± 0.007**	**0.749 ± 0.010**	0.746 ± 0.010	0.746 ± 0.010
I [10^5] ↑	Uniform (baseline)	1.034 ± 0.052	2.923 ± 0.122	3.485 ± 0.059	3.505 ± 0.061
	All-ones	2.599 ± 0.125	3.398 ± 0.074	3.532 ± 0.064	3.532 ± 0.066
	StartGrad (ours)	**3.538 ± 0.074**	**3.538 ± 0.060**	3.545 ± 0.059	3.544 ± 0.059
E [10^4] ↓	Uniform (baseline)	1.914 ± 0.137	0.921 ± 0.060	0.735 ± 0.027	0.729 ± 0.027
	All-ones	1.473 ± 0.075	1.047 ± 0.037	0.729 ± 0.027	0.726 ± 0.026
	StartGrad (ours)	**1.232 ± 0.055**	**0.753 ± 0.033**	0.720 ± 0.027	0.720 ± 0.027

Table 6: Average performance for the ExtremalMask (Enguehard, 2023) explanation method (deletion game objective) across iteration steps for the state dataset when initialized with StartGrad, all-ones and uniformly. The reported numbers are the mean and standard deviation across five folds. For each metric, ↑ indicates that higher is better, and ↓ that lower is better. Baseline refers to the initialization scheme originally used for the respective mask explanation method. Outcomes that are one standard deviation away from the second-best one are highlighted in bold.

		Iteration steps			
Metric	Initialization	50	100	300	500
AUP ↑	Uniform (baseline)	0.630 ± 0.029	0.712 ± 0.024	0.821 ± 0.028	0.870 ± 0.026
	All-ones	0.824 ± 0.017	**0.895 ± 0.017**	**0.909 ± 0.022**	**0.916 ± 0.022**
	StartGrad (ours)	**0.843 ± 0.018**	**0.895 ± 0.024**	**0.917 ± 0.025**	**0.925 ± 0.025**
AUR ↑	Uniform (baseline)	0.673 ± 0.012	0.717 ± 0.011	0.754 ± 0.011	0.752 ± 0.009
	All-ones	0.481 ± 0.020	0.600 ± 0.011	0.747 ± 0.011	0.752 ± 0.010
	StartGrad (ours)	**0.815 ± 0.004**	**0.763 ± 0.007**	0.754 ± 0.009	0.754 ± 0.009
I [10^5] ↑	Uniform (baseline)	1.072 ± 0.049	3.948 ± 0.064	4.326 ± 0.032	4.377 ± 0.039
	All-ones	0.739 ± 0.052	0.891 ± 0.030	4.340 ± 0.042	4.395 ± 0.042
	StartGrad (ours)	**4.415 ± 0.037**	**4.416 ± 0.042**	**4.416 ± 0.041**	4.413 ± 0.040
E [10^4] ↓	Uniform (baseline)	1.536 ± 0.053	0.479 ± 0.019	0.348 ± 0.048	0.283 ± 0.037
	All-ones	1.632 ± 0.119	1.875 ± 0.135	**0.287 ± 0.036**	0.264 ± 0.341
	StartGrad (ours)	**0.932 ± 0.016**	**0.350 ± 0.015**	**0.261 ± 0.031**	0.260 ± 0.032

E.2.2 Switch-Feature Dataset

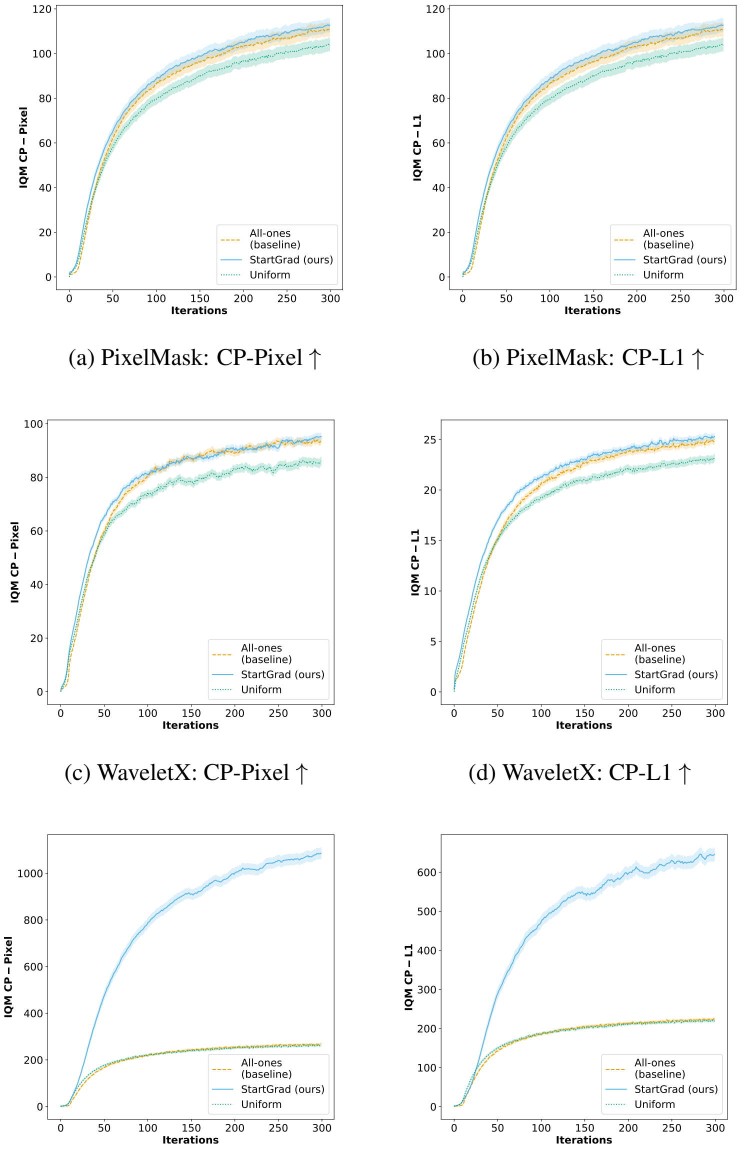
Table 7: Average performance for the ExtremalMask (Enguehard, 2023) explanation method (deletion game objective) across iteration steps for the switch-feature dataset when initialized with StartGrad, all-ones and uniformly. The reported numbers are the mean and standard deviation across five folds. For each metric, ↑ indicates that higher is better, and ↓ that lower is better. Baseline refers to the initialization scheme originally used for the respective mask explanation method. Outcomes that are one standard deviation away from the second-best one are highlighted in bold.

		Iteration steps			
Metric	Initialization	50	100	300	500
AUP ↑	Uniform (baseline)	0.825 ± 0.025	0.875 ± 0.041	**0.913 ± 0.071**	**0.932 ± 0.078**
	All-ones	0.762 ± 0.016	0.705 ± 0.047	0.764 ± 0.013	0.755 ± 0.015
	StartGrad (ours)	**0.869 ± 0.007**	**0.908 ± 0.019**	**0.936 ± 0.048**	**0.948 ± 0.048**
AUR ↑	Uniform (baseline)	0.629 ± 0.017	**0.622 ± 0.017**	**0.631 ± 0.027**	**0.629 ± 0.029**
	All-ones	**0.695 ± 0.038**	0.545 ± 0.026	0.563 ± 0.012	0.547 ± 0.013
	StartGrad (ours)	**0.693 ± 0.011**	**0.634 ± 0.013**	**0.629 ± 0.024**	**0.628 ± 0.029**
I $[10^5]$ ↑	Uniform (baseline)	0.781 ± 0.052	1.532 ± 0.112	**2.232 ± 0.217**	**2.248 ± 0.241**
	All-ones	0.923 ± 0.128	0.559 ± 0.053	1.410 ± 0.037	1.548 ± 0.058
	StartGrad (ours)	**1.734 ± 0.103**	**2.046 ± 0.072**	**2.211 ± 0.155**	**2.220 ± 0.168**
E $[10^4]$ ↓	Uniform (baseline)	2.639 ± 0.163	1.842 ± 0.077	**1.410 ± 0.082**	**1.378 ± 0.074**
	All-ones	**2.411 ± 0.210**	3.006 ± 0.072	1.985 ± 0.053	1.717 ± 0.051
	StartGrad (ours)	**2.399 ± 0.044**	**1.680 ± 0.065**	**1.398 ± 0.073**	**1.366 ± 0.060**

F Additional Ablation Studies

F.1 Vision Experiments with VGG16 as baseline model

F.1.1 Figures for the VGG16 experiment



(a) PixelMask: CP-Pixel ↑

(b) PixelMask: CP-L1 ↑

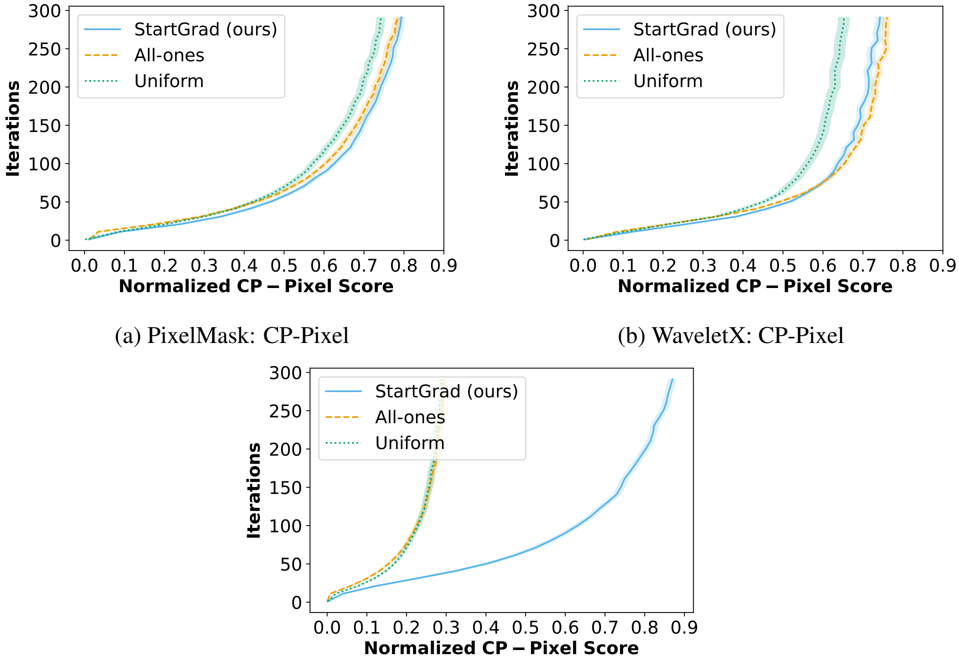
(c) WaveletX: CP-Pixel ↑

(d) WaveletX: CP-L1 ↑

(e) ShearletX: CP-Pixel ↑

(f) ShearletX: CP-L1 ↑

Figure 10: Comparison of StartGrad initialization (ours) with standard baseline initialization schemes for the PixelMask (Fong & Vedaldi, 2017) (first row), WaveletX (Kolek et al., 2023) (second row) and ShearletX (Kolek et al., 2023) (third row) explanation models. Baseline refers to the originally used initialization scheme for the respective mask explanation method. The solid line represents the average of the interquartile mean (IQM) performance across 500 randomly selected validation ImageNet (Deng et al., 2009) samples, while the shaded area denotes the standard errors respectively. For each metric, ↑ indicates that higher is better, and ↓ that lower is better. We employ a pretrained VGG16 (Simonyan & Zisserman, 2014) model as a classifier.

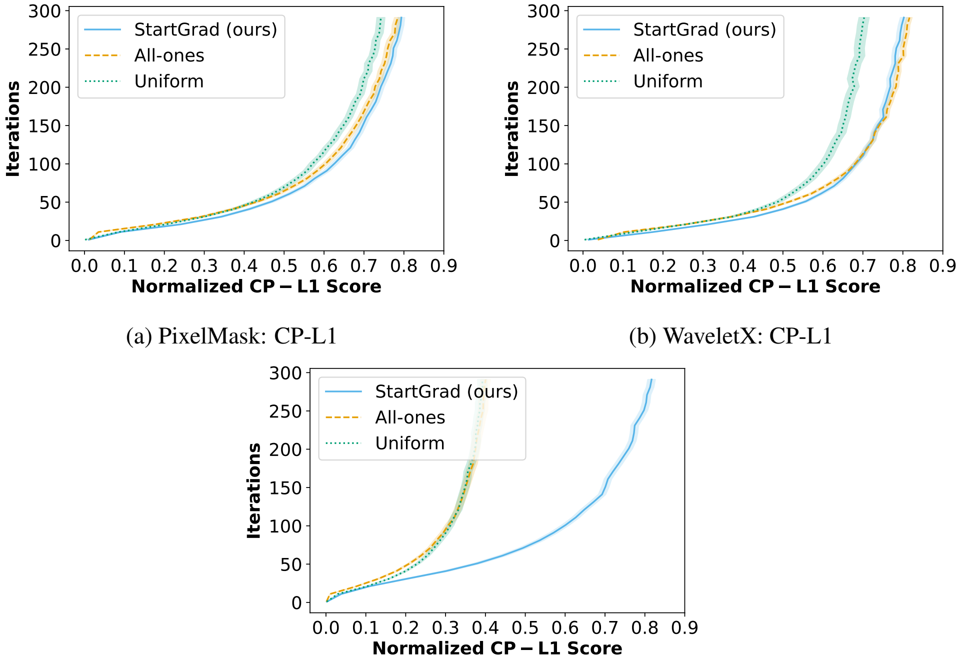


(a) PixelMask: CP-Pixel

(b) WaveletX: CP-Pixel

(c) ShearletX: CP-Pixel

Figure 11: Normalized CP-Pixel score vs. iteration steps for the PixelMask (Fong & Vedaldi, 2017) (a), WaveletX (Kolek et al., 2023) (b), and ShearletX (Kolek et al., 2023) (c) models, comparing three initialization methods: StartGrad (ours) (solid line), All-ones (dashed line), and Uniform (dotted line). The curves represent the average number of iteration steps needed to reach a normalized target CP-L1 score. This target score is defined as the highest CP-L1 value achieved across all three initialization methods for each model, then normalized by the overall maximum score observed. The shaded regions represent the standard error across 500 randomly selected ImageNet validation samples (Deng et al., 2009), with a pretrained VGG16 (Simonyan & Zisserman, 2014) classifier. StartGrad allows all three models to reach target scores in fewer iterations than the uniform initialization. Compared to all-ones initialization, it accelerates PixelMask (Fong & Vedaldi, 2017) and ShearletX (Kolek et al., 2023) while performing slightly worse for WaveletX (Kolek et al., 2023) for medium to high scores.



(a) PixelMask: CP-L1

(b) WaveletX: CP-L1

(c) ShearletX: CP-L1

Figure 12: Normalized CP-L1 score vs. iteration steps for the PixelMask (Fong & Vedaldi, 2017) (a), WaveletX (Kolek et al., 2023) (b), and ShearletX (Kolek et al., 2023) (c) models, comparing three initialization methods: StartGrad (ours) (solid line), All-ones (dashed line), and Uniform (dotted line). The curves represent the average number of iteration steps needed to reach a normalized target CP-L1 score. This target score is defined as the highest CP-L1 value achieved across all three initialization methods for each model, then normalized by the overall maximum score observed. The shaded regions represent the standard error across 500 randomly selected ImageNet validation samples (Deng et al., 2009), with a pretrained VGG16 (Simonyan & Zisserman, 2014) classifier. StartGrad allows all three models to reach target scores in fewer iterations than the uniform initialization. Compared to all-ones initialization, it accelerates PixelMask (Fong & Vedaldi, 2017) and ShearletX (Kolek et al., 2023) while performing slightly worse for WaveletX (Kolek et al., 2023) for medium to high scores.

F.2 Comparison Performance Differences Startgrad vs. baselines

Table 8: Median pairwise performance difference between StartGrad and baseline initialization methods across different iteration steps for 500 randomly selected validation ImageNet (Deng et al., 2009) samples using a VGG16 (Simonyan & Zisserman, 2014) classifier. Baseline refers to the originally used initialization scheme for the respective mask explanation method. Statistical significance is denoted by **, **, and * at the 1%, 5%, and 10% levels, respectively, based on an one-sided Wilcoxon signed-rank test with Bonferroni correction at the method and metric level. Since PixelMask (Fong & Vedaldi, 2017) does only apply a mask to the pixel space, the CP-Pixel and CP-L1 scores are identical.

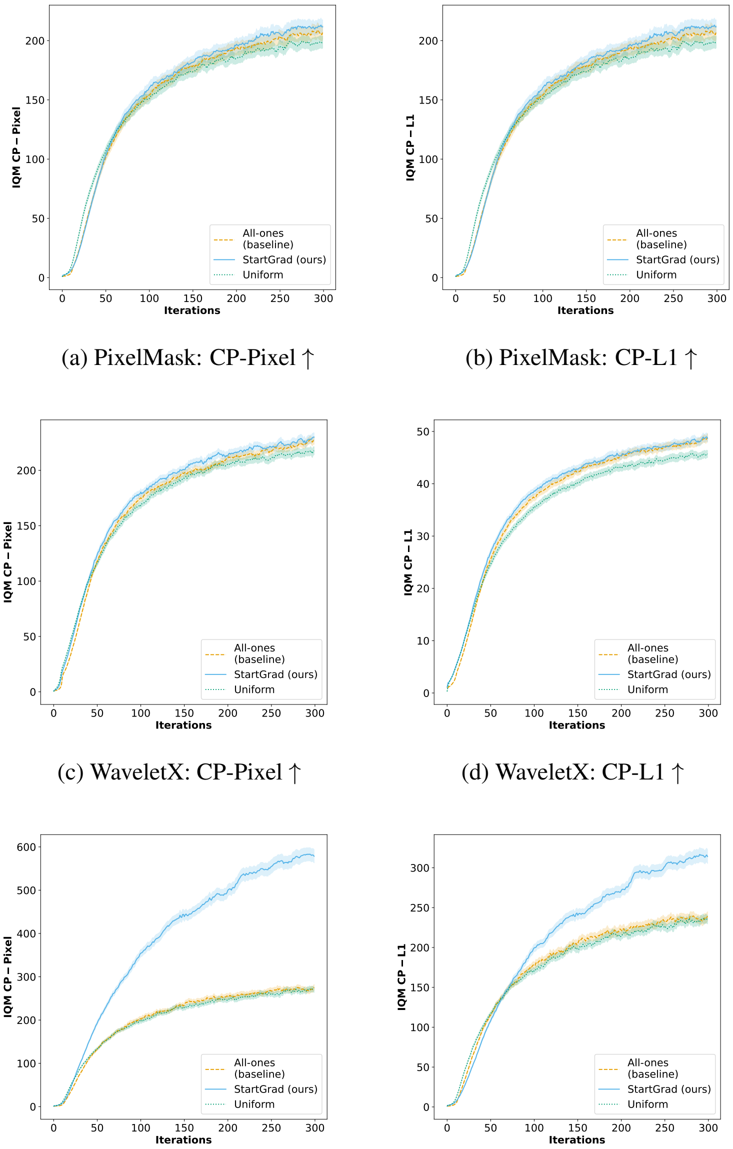
		$\triangle$ CP-Pixel ↑			$\triangle$ CP-L1 ↑		
		Iteration steps			Iteration steps		
Method	Baseline initialization	50	100	300	50	100	300
PixelMask	All-ones (baseline)	2.29^{***}	2.04^{***}	1.65^{**}	2.29^{***}	2.04^{***}	1.65^{**}
	Uniform	2.60^{***}	4.06^{***}	3.33^{***}	2.60^{***}	4.06^{***}	3.33^{***}
WaveletX	All-ones (baseline)	3.61^{***}	0.88	-0.11	1.11^{***}	0.52^{**}	0.10
	Uniform	2.56^{***}	2.04^{***}	2.82^{***}	0.85^{***}	0.67^{***}	0.85^{***}
ShearletX	All-ones (baseline)	180.31^{***}	454.65^{***}	767.72^{***}	55.22^{***}	192.11^{***}	351.26^{***}
	Uniform	173.49^{***}	462.95^{***}	770.24^{***}	52.93^{***}	201.90^{***}	370.77^{***}

Table 9: Average iteration steps required for the ShearletX (Kolek et al., 2023) method initialized with StartGrad to match the corresponding performance of the same method initialized with all-ones or uniform. The table reports the time (in iteration steps) needed to match the interquartile mean (IQM) and median performance of the CP-Pixel and CP-L1 score obtained. Additionally, speedup is calculated as the ratio between the baseline iteration steps and the time taken under StartGrad initialization. All experiments use 500 random ImageNet (Deng et al., 2009) samples evaluated on a pretrained VGG16 (Simonyan & Zisserman, 2014) classifier. The target metrics are obtained for running the ShearletX method across 300 iteration steps. To obtain the average and standard error, we use bootstrapping across 250 resampled sets. For each metric, ↑ indicates that higher is better, and ↓ that lower is better. Values that are 2 standard errors away from the baseline values (300 for iteration, 1 for speedup) are highlighted in bold.

Reference	Target metric (at 300 iterations)	CP-Pixel		CP-L1	
		Iterations ↓	Speedup ↑	Iterations ↓	Speedup ↑
All-ones (baseline)	IQM	**33.10 ± 1.55**	**9.05 ± 0.42**	**41.43 ± 2.30**	**7.24 ± 0.40**
	Median	**36.75 ± 2.57**	**8.18 ± 0.58**	**48.82 ± 4.08**	**6.17 ± 0.52**
Uniform	IQM	**32.58 ± 1.50**	**9.20 ± 0.42**	**40.79 ± 2.23**	**7.35 ± 0.40**
	Median	**36.23 ± 2.49**	**8.29 ± 0.58**	**48.20 ± 4.02**	**6.25 ± 0.53**

F.3 Vision Experiments with Swin Transformer as baseline model

F.3.1 Figures for the Swin Transformer experiment



(a) PixelMask: CP-Pixel ↑

(b) PixelMask: CP-L1 ↑

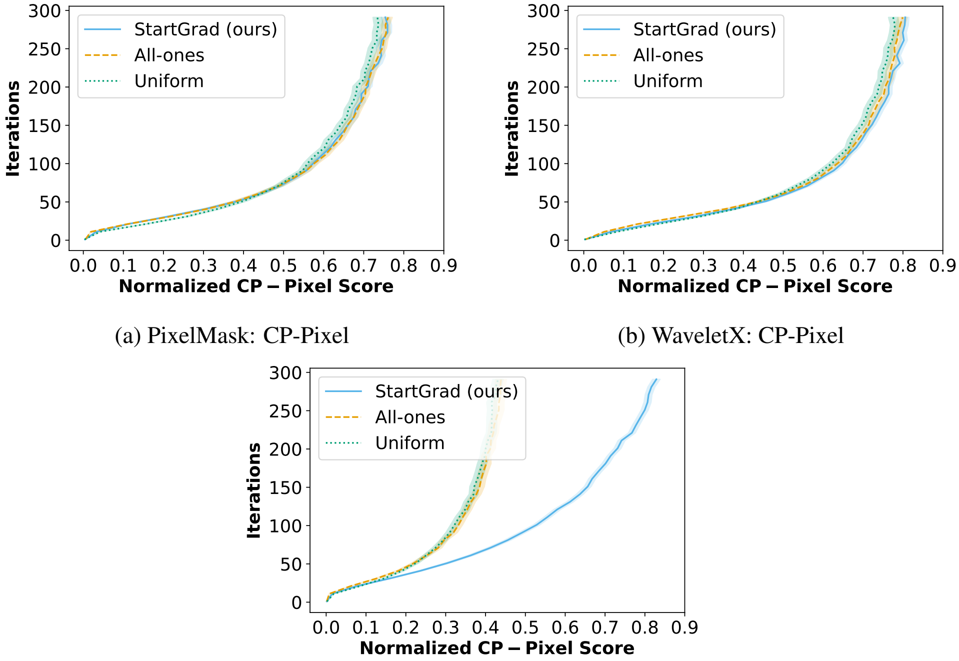
(c) WaveletX: CP-Pixel ↑

(d) WaveletX: CP-L1 ↑

(e) ShearletX: CP-Pixel ↑

(f) ShearletX: CP-L1 ↑

Figure 13: Comparison of StartGrad initialization (ours) with standard baseline initialization schemes for the PixelMask (Fong & Vedaldi, 2017) (first row), WaveletX (Kolek et al., 2023) (second row) and ShearletX (Kolek et al., 2023) (third row) explanation models. Baseline refers to the originally used initialization scheme for the respective mask explanation method. The solid line represents the average of the interquartile mean (IQM) performance across 250 randomly selected validation ImageNet (Deng et al., 2009) samples, while the shaded area denotes the standard errors respectively. For each metric, ↑ indicates that higher is better, and ↓ that lower is better. We employ a pretrained Swin Transformer (Liu et al., 2022) model as a classifier.

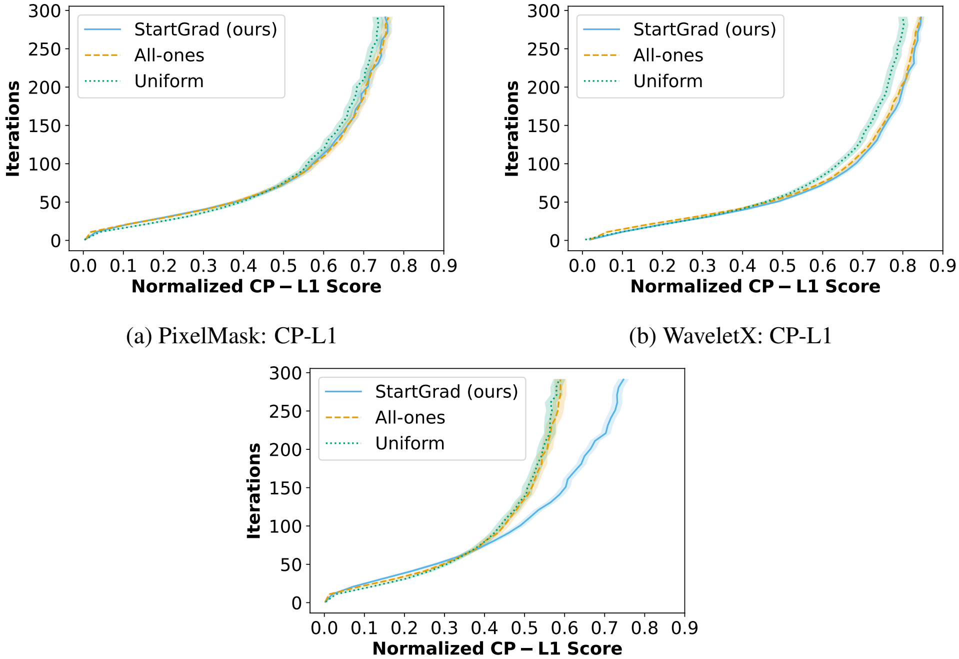


(a) PixelMask: CP-Pixel

(b) WaveletX: CP-Pixel

(c) ShearletX: CP-Pixel

Figure 14: Normalized CP-Pixel score vs. iteration steps for the PixelMask (Fong & Vedaldi, 2017) (a), WaveletX (Kolek et al., 2023) (b), and ShearletX (Kolek et al., 2023) (c) models, comparing three initialization methods: StartGrad (ours) (solid line), All-ones (dashed line), and Uniform (dotted line). The curves represent the average number of iteration steps needed to reach a normalized target CP-Pixel score. This target score is defined as the highest CP-Pixel value achieved across all three initialization methods for each model, then normalized by the overall maximum score observed. The shaded regions represent the standard error across 250 randomly selected ImageNet validation samples (Deng et al., 2009), with a pretrained Swin Transformer (Liu et al., 2022) classifier. StartGrad enables all three models to achieve target scores with fewer iterations compared to uniform initialization scheme. Compared to the all-ones initialization, StartGrad enables to achieve target scores with fewer iterations for the WaveletX (Kolek et al., 2023) and ShearletX (Kolek et al., 2023) model, and performs on par for the PixelMask (Fong & Vedaldi, 2017) model.



(a) PixelMask: CP-L1

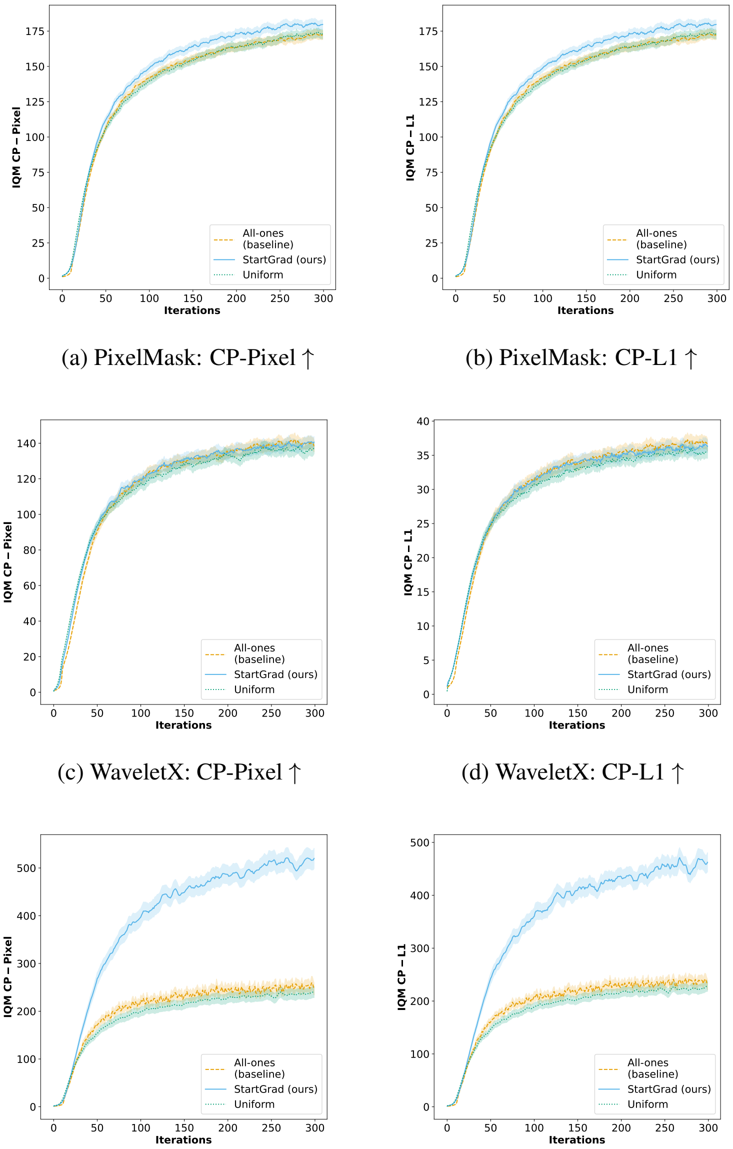
(b) WaveletX: CP-L1

(c) ShearletX: CP-L1

Figure 15: Normalized CP-L1 score vs. iteration steps for the PixelMask (Fong & Vedaldi, 2017) (a), WaveletX (Kolek et al., 2023) (b), and ShearletX (Kolek et al., 2023) (c) models, comparing three initialization methods: StartGrad (ours) (solid line), All-ones (dashed line), and Uniform (dotted line). The curves represent the average number of iteration steps needed to reach a normalized target CP-L1 score. This target score is defined as the highest CP-L1 value achieved across all three initialization methods for each model, then normalized by the overall maximum score observed. The shaded regions represent the standard error across 250 randomly selected ImageNet validation samples (Deng et al., 2009), with a pretrained Swin Transformer (Liu et al., 2022) classifier. StartGrad enables all three models to achieve target scores with fewer iterations compared to uniform initialization scheme. Compared to the all-ones initialization, StartGrad enables to achieve target scores with fewer iterations for the WaveletX (Kolek et al., 2023) and ShearletX (Kolek et al., 2023) model, and performs on par for the PixelMask (Fong & Vedaldi, 2017) model.

F.4 Vision Experiments with Vision Transformer as baseline model

F.4.1 Figures for the Vision Transformer experiment



(a) PixelMask: CP-Pixel ↑

(b) PixelMask: CP-L1 ↑

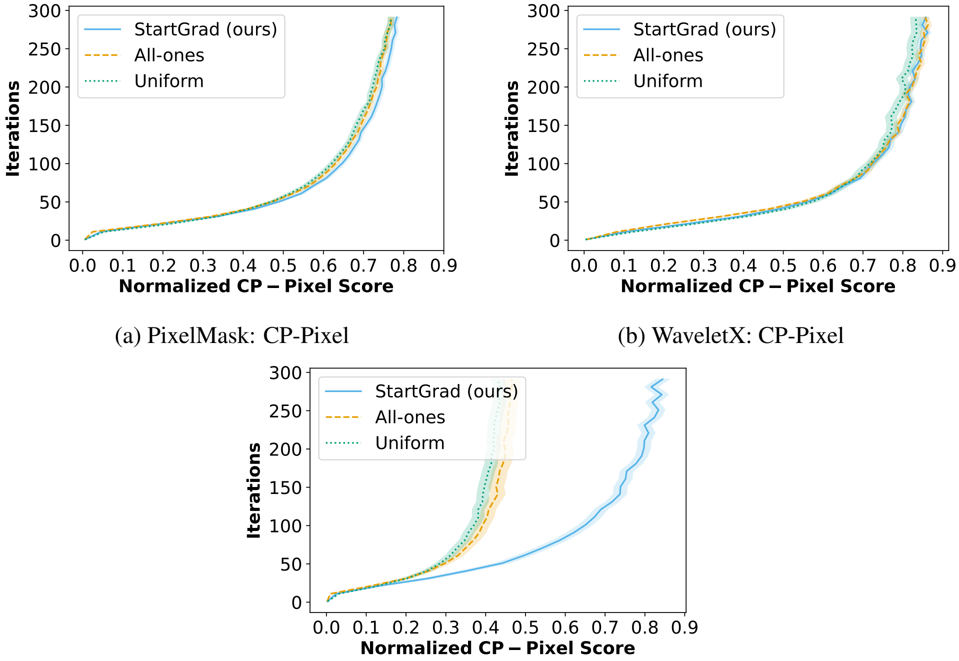
(c) WaveletX: CP-Pixel ↑

(d) WaveletX: CP-L1 ↑

(e) ShearletX: CP-Pixel ↑

(f) ShearletX: CP-L1 ↑

Figure 16: Comparison of StartGrad initialization (ours) with standard baseline initialization schemes for the PixelMask (Fong & Vedaldi, 2017) (first row), WaveletX (Kolek et al., 2023) (second row) and ShearletX (Kolek et al., 2023) (third row) explanation models. Baseline refers to the originally used initialization scheme for the respective mask explanation method. The solid line represents the average of the interquartile mean (IQM) performance across 100 randomly selected validation ImageNet (Deng et al., 2009) samples, while the shaded area denotes the standard errors respectively. For each metric, ↑ indicates that higher is better, and ↓ that lower is better. We employ a pretrained Vision Transformer model (Dosovitskiy et al., 2021) model as a classifier.

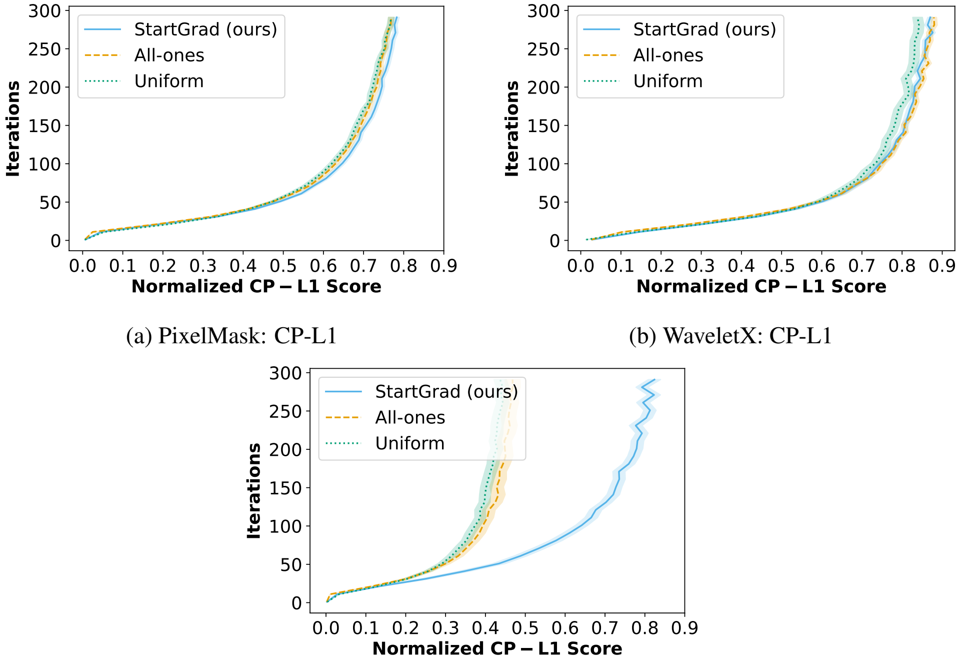


(a) PixelMask: CP-Pixel

(b) WaveletX: CP-Pixel

(c) ShearletX: CP-Pixel

Figure 17: Normalized CP-Pixel score vs. iteration steps for the PixelMask (Fong & Vedaldi, 2017) (a), WaveletX (Kolek et al., 2023) (b), and ShearletX (Kolek et al., 2023) (c) models, comparing three initialization methods: StartGrad (ours) (solid line), All-ones (dashed line), and Uniform (dotted line). The curves represent the average number of iteration steps needed to reach a normalized target CP-Pixel score. This target score is defined as the highest CP-Pixel value achieved across all three initialization methods for each model, then normalized by the overall maximum score observed. The shaded regions represent the standard error across 100 randomly selected ImageNet validation samples (Deng et al., 2009), with a Vision Transformer model (Dosovitskiy et al., 2021) classifier. StartGrad enables all three models to achieve target scores with fewer iterations compared to uniform initialization scheme. Compared to the all-ones initialization, StartGrad enables to achieve target scores with fewer iterations for the PixelMask (Fong & Vedaldi, 2017) and ShearletX (Kolek et al., 2023) model, and performs on par for the WaveletX (Kolek et al., 2023) model.



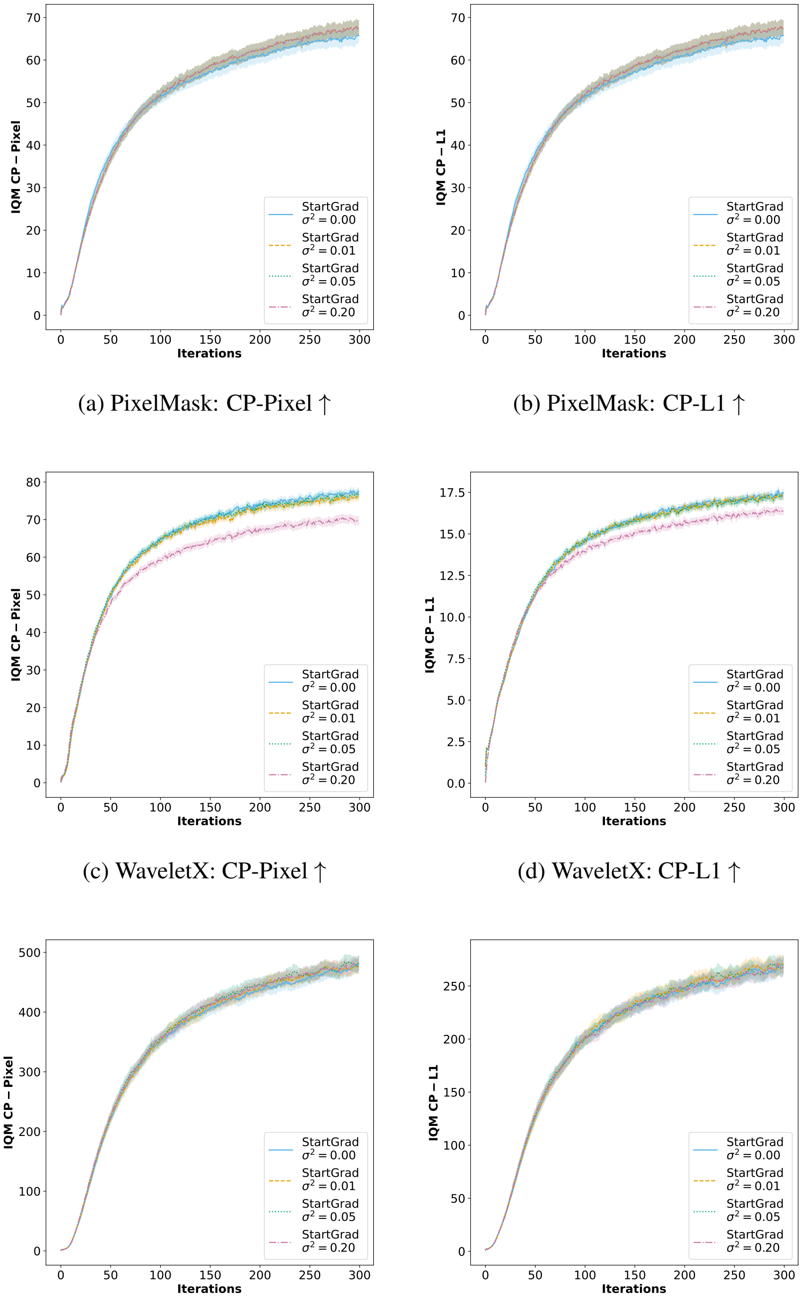
(a) PixelMask: CP-L1

(b) WaveletX: CP-L1

(c) ShearletX: CP-L1

Figure 18: Normalized CP-L1 score vs. iteration steps for the PixelMask (Fong & Vedaldi, 2017) (a), WaveletX (Kolek et al., 2023) (b), and ShearletX (Kolek et al., 2023) (c) models, comparing three initialization methods: StartGrad (ours) (solid line), All-ones (dashed line), and Uniform (dotted line). The curves represent the average number of iteration steps needed to reach a normalized target CP-L1 score. This target score is defined as the highest CP-L1 value achieved across all three initialization methods for each model, then normalized by the overall maximum score observed. The shaded regions represent the standard error across 100 random ImageNet validation samples (Deng et al., 2009), with a Vision Transformer model (Dosovitskiy et al., 2021) classifier. StartGrad allows all three models to reach target scores in fewer iterations than the uniform initialization. Compared to all-ones initialization, it accelerates PixelMask (Fong & Vedaldi, 2017) and ShearletX (Kolek et al., 2023) while performing on par with WaveletX (Kolek et al., 2023).

F.5 Effect of noisy gradients on StartGrad



(a) PixelMask: CP-Pixel ↑

(b) PixelMask: CP-L1 ↑

(c) WaveletX: CP-Pixel ↑

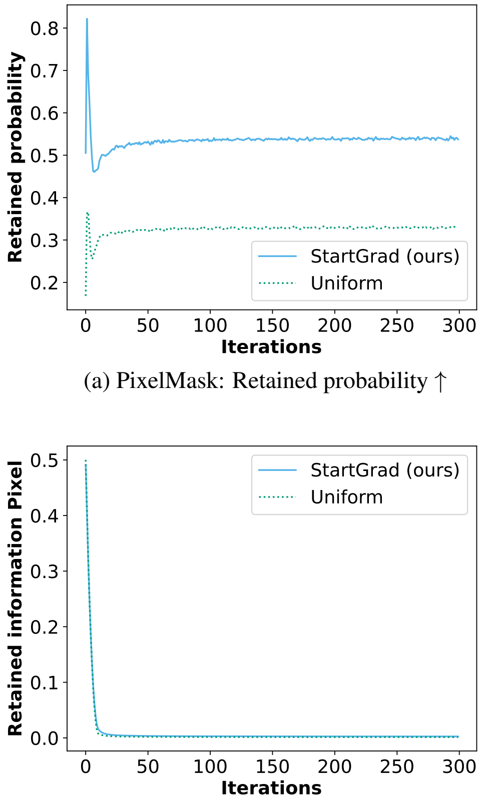
(d) WaveletX: CP-L1 ↑

(e) ShearletX: CP-Pixel ↑

(f) ShearletX: CP-L1 ↑

Figure 19: Investigation of StartGrad's robustness to noisy gradients by progressively multiplying the gradient with Gaussian noise of increasing standard deviation for the PixelMask (Fong & Vedaldi, 2017) (first row), WaveletX (Kolek et al., 2023) (second row) and ShearletX (Kolek et al., 2023) (third row) explanation models. The solid line represents the average of the interquartile mean (IQM) performance across 500 randomly selected validation ImageNet (Deng et al., 2009) samples, while the shaded area denotes the standard errors respectively. ↑ indicates that higher is better, and ↓ that lower is better. We use a pretrained ResNet18 (He et al., 2016) classifier.

F.6 Impact of Gradient-Based Initialization on Distortion and Sparsity Dynamics for PixelMask

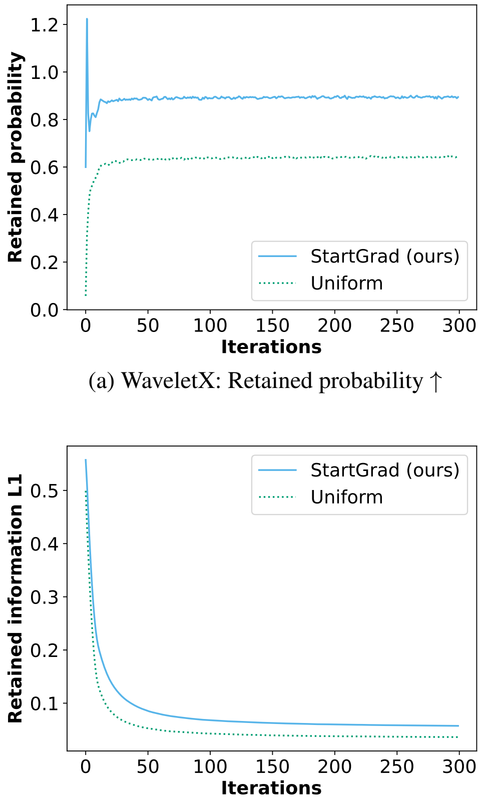


(a) PixelMask: Retained probability ↑

(b) PixelMask: Retained information Pixel ↓

Figure 20: Overview of optimization dynamics of retained probability (a), measured as the ratio between the masked input prediction and the original prediction, and retained information (in pixel space) (b), for the StartGrad (ours) and uniform initialization method, using the PixelMask (Fong & Vedaldi, 2017) model. We employ a pretrained ResNet18 (He et al., 2016) model as a classifier. The solid lines represent the average performance across 500 randomly selected ImageNet validation samples (Deng et al., 2009). For each metric, ↑ indicates that higher is better, and ↓ that lower is better. Results indicate that the gradient-information that StartGrad uses to initialize the mask guides the mask-based explanation method effectively to achieve minimal distortion (as indicated by high retained probability) early in the optimization process, consistent with theoretical predictions. However, this comes at the cost of reduced sparsity, as evidenced by higher retained information (b). In contrast, uniform initialization exhibits better sparsity (lower retained information) but struggles to retain critical information as indicated by comparable low retained probability, highlighting a trade-off between higher retained probability and increased sparsity. The optimization dynamics further reveal that the early iteration steps are especially crucial for retaining probability, as the mask-based explanation methods quickly converge in terms of retained probability for both initialization schemes. This behavior underscores the importance of leveraging gradient signals for efficient initialization, achieving a favorable starting point for optimization in terms of minimal distortion (high retained probability).

F.7 Impact of Gradient-Based Initialization on Distortion and Sparsity Dynamics for WaveletX

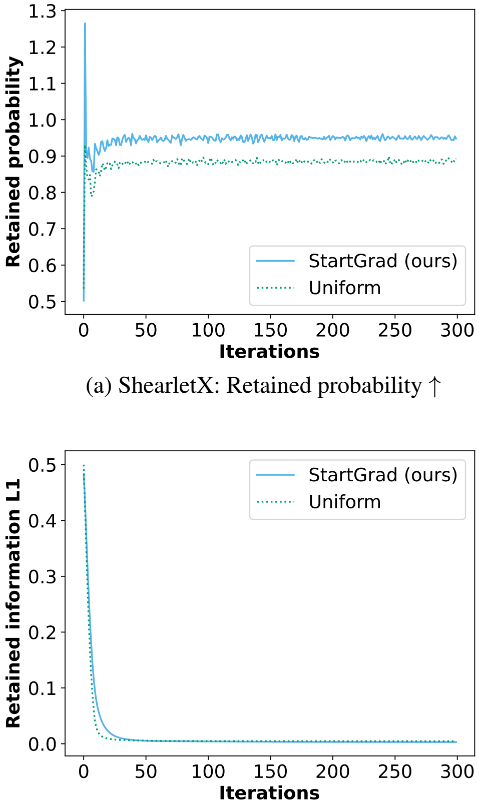


(a) WaveletX: Retained probability ↑

(b) WaveletX: Retained information L1 ↓

Figure 21: Overview of optimization dynamics of retained probability (a), measured as the ratio between the masked input prediction and the original prediction, and retained information (L_1) (b), for the StartGrad (ours) and uniform initialization method, using the WaveletX model (Kolek et al., 2023). We employ a pretrained ResNet18 (He et al., 2016) model as a classifier. The solid lines represent the average performance across 500 randomly ImageNet validation (Deng et al., 2009) samples. For each metric, ↑ indicates that higher is better, and ↓ that lower is better. Results indicate that the gradient-information that StartGrad uses to initialize the mask guides the mask-based explanation method effectively to achieve minimal distortion (as indicated by high retained probability) early in the optimization process, consistent with theoretical predictions. However, this comes at the cost of reduced sparsity, as evidenced by higher retained information (b). In contrast, uniform initialization exhibits better sparsity (lower retained information) but struggles to retain critical information as indicated by comparable low retained probability, highlighting a trade-off between higher retained probability and increased sparsity. The optimization dynamics further reveal that the early iteration steps are especially crucial for retaining probability, as the mask-based explanation methods quickly converge in terms of retained probability for both initialization schemes. This behavior underscores the importance of leveraging gradient signals for efficient initialization, achieving a favorable starting point for optimization in terms of minimal distortion (high retained probability).

F.8 Impact of Gradient-Based Initialization on Distortion and Sparsity Dynamics for ShearletX

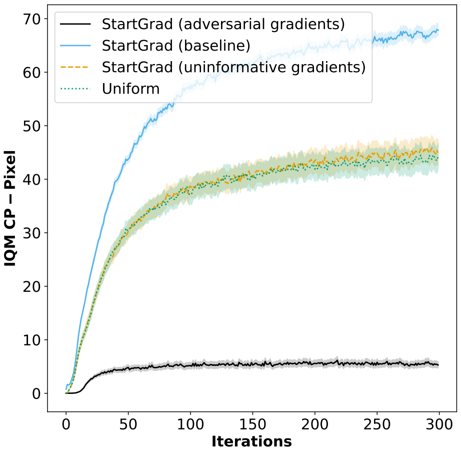


(a) ShearletX: Retained probability ↑

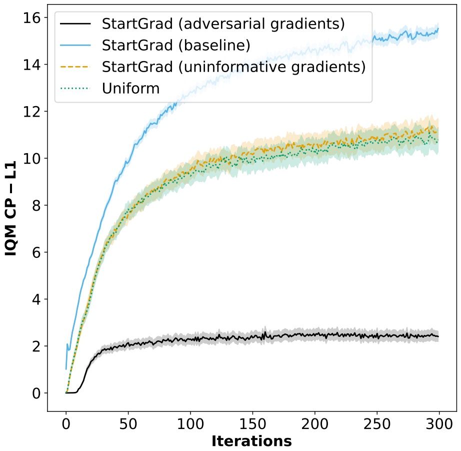
(b) ShearletX: Retained information L1 ↓

Figure 22: Overview of optimization dynamics of retained probability (a), measured as the ratio between the masked input prediction and the original prediction, and retained information (L_1) (b), for the StartGrad (ours) and uniform initialization method, using the ShearletX model (Kolek et al., 2023). We employ a pretrained ResNet18 (He et al., 2016) model as a classifier. The solid lines represent the average performance across 500 randomly selected ImageNet (Deng et al., 2009) validation samples. For each metric, ↑ indicates that higher is better, and ↓ that lower is better. Results indicate that the gradient-information that StartGrad uses to initialize the mask guides the mask-based explanation method effectively to achieve minimal distortion (as indicated by high retained probability) early in the optimization process, consistent with theoretical predictions. However, this comes at the cost of reduced sparsity, as evidenced by higher retained information (b). In contrast, uniform initialization exhibits better sparsity (lower retained information) but struggles to retain critical information as indicated by comparable low retained probability, highlighting a trade-off between higher retained probability and increased sparsity. The optimization dynamics further reveal that the early iteration steps are especially crucial for retaining probability, as the mask-based explanation methods quickly converge in terms of retained probability for both initialization schemes. This behavior underscores the importance of leveraging gradient signals for efficient initialization, achieving a favorable starting point for optimization in terms of minimal distortion (high retained probability).

F.9 Effect of uninformative and adversarial gradients on StartGrad



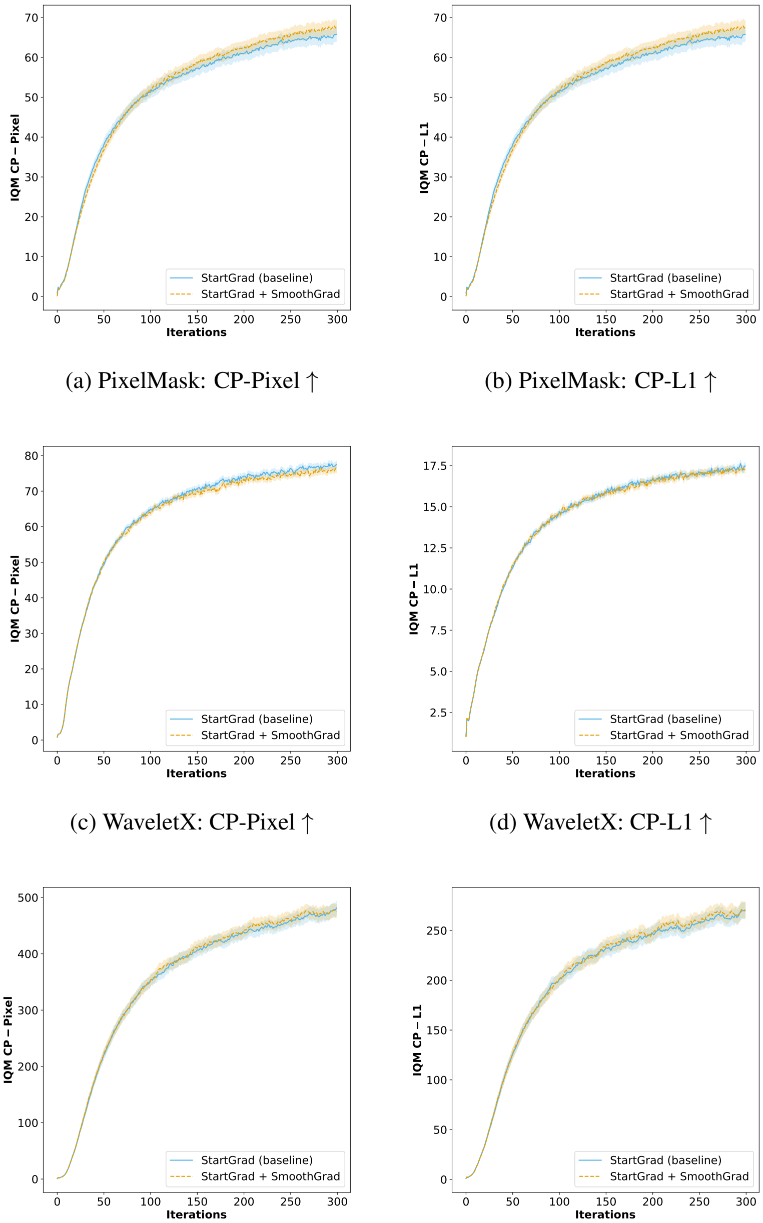
(a) WaveletX: CP-Pixel ↑



(b) WaveletX: CP-L1 ↑

Figure 23: Investigation of the robustness of StartGrad initialization under scenarios where gradients are either uninformative or adversarial (misleading) for the WaveletX model (Kolek et al., 2023). In the uninformative setting, gradient values were randomly shuffled to destroy their structural signal, effectively rendering them meaningless for guiding mask initialization. In the adversarial setting, the correspondence between gradient values and mask initialization values was reversed, such that features with higher gradient values were assigned lower mask values and vice versa.The solid line represents the average of the interquartile mean (IQM) performance across 250 randomly selected validation ImageNet (Deng et al., 2009) samples, while the shaded area denotes the standard errors respectively. For each metric, ↑ indicates that higher is better, and ↓ that lower is better. We employ a pretrained ResNet18 (He et al., 2016) model as a classifier. Results reveal that under uninformative gradient conditions, StartGrad's performance is comparable to the baseline uniform initialization, indicating that StartGrad relies on meaningful gradient signals to offer an advantage. In the adversarial setting, performance degradation highlights the importance of smart initialization, as the algorithm struggles to recover from a poorly initialized mask during optimization. These findings emphasize the critical role of robust and informed initialization in achieving optimal performance.

F.10 Effect of using different gradient methods on StartGrad



(a) PixelMask: CP-Pixel ↑

(b) PixelMask: CP-L1 ↑

(c) WaveletX: CP-Pixel ↑

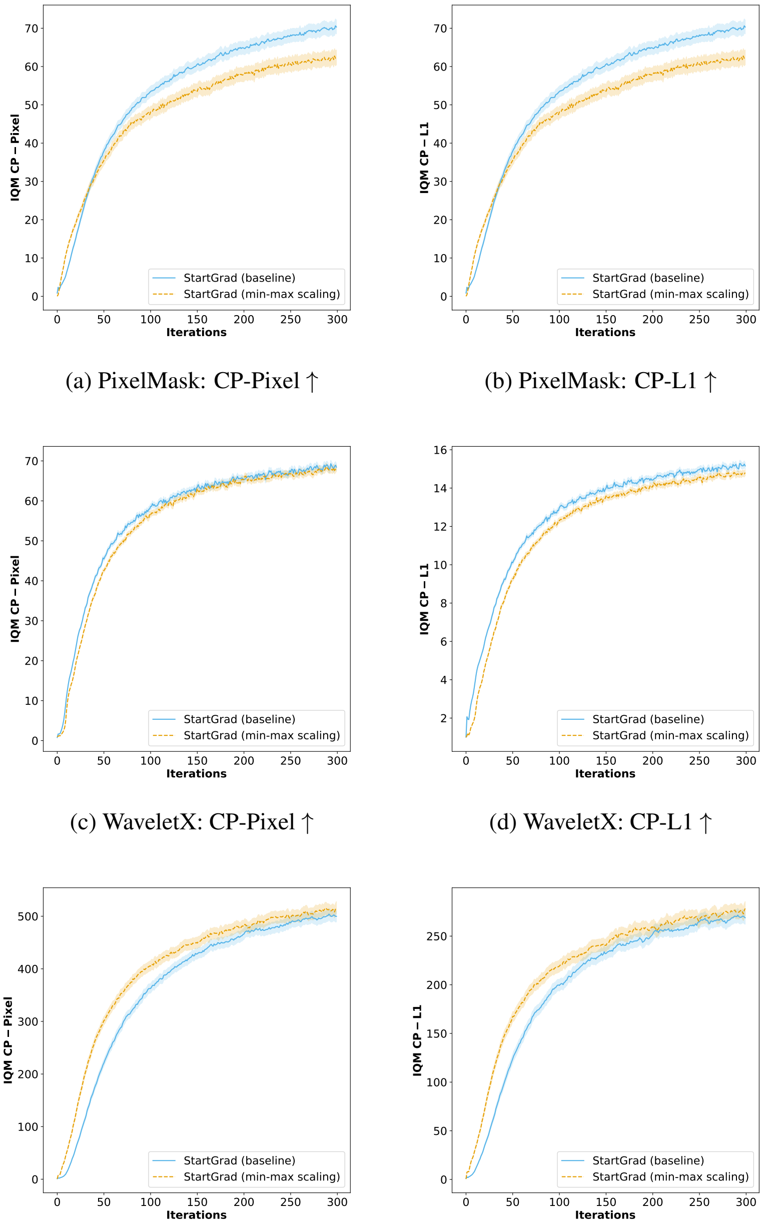
(d) WaveletX: CP-L1 ↑

(e) ShearletX: CP-Pixel ↑

(f) ShearletX: CP-L1 ↑

Figure 24: Effect of using a more advanced gradient-based saliency method (here SmoothGrad (Smilkov et al., 2017)) in connection with our proposed StartGrad algorithm for the PixelMask (Fong & Vedaldi, 2017) (first row), WaveletX (Kolek et al., 2023) (second row) and ShearletX (Kolek et al., 2023) (third row) explanation models. The solid line represents the average of the interquartile mean (IQM) performance across 500 randomly selected validation ImageNet (Deng et al., 2009) samples, while the shaded area denotes the standard errors respectively. For each metric, ↑ indicates that higher is better, and ↓ that lower is better. We employ a pretrained ResNet18 (He et al., 2016) model as a classifier.

F.11 Effect of using a different transformation function instead of QTF



(a) PixelMask: CP-Pixel ↑

(b) PixelMask: CP-L1 ↑

(c) WaveletX: CP-Pixel ↑

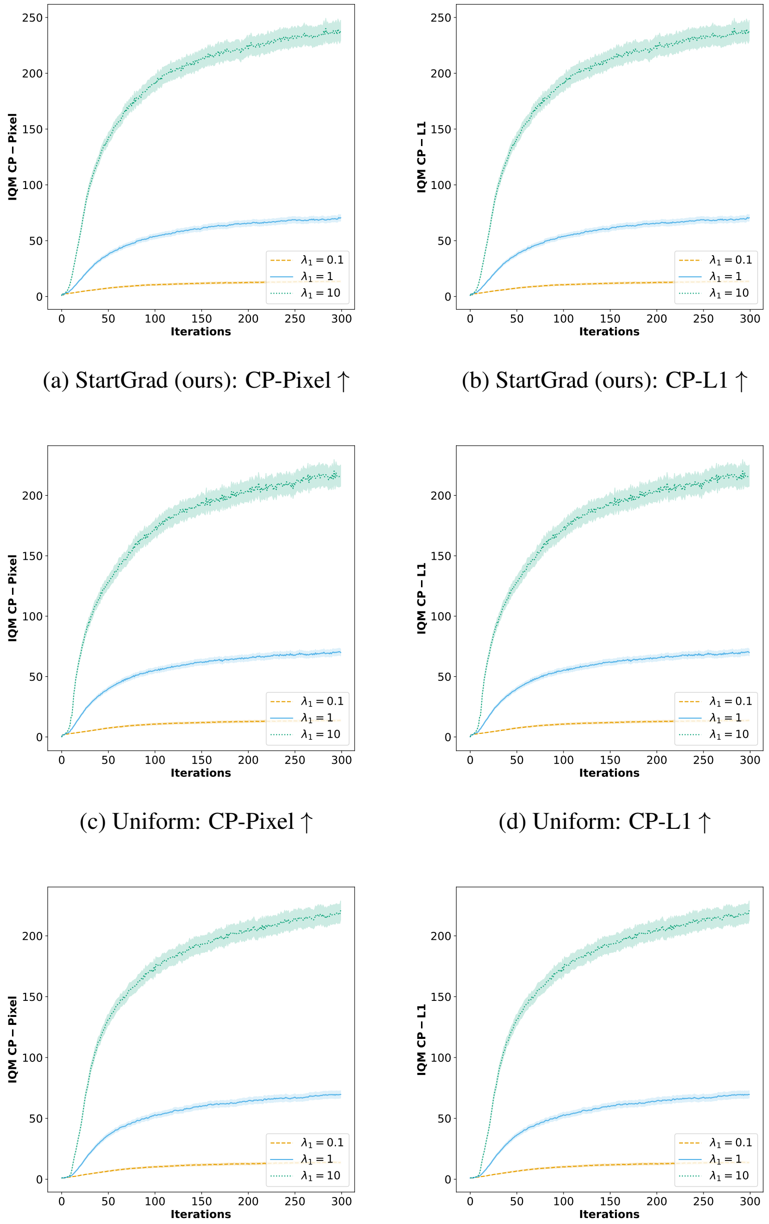
(d) WaveletX: CP-L1 ↑

(e) ShearletX: CP-Pixel ↑

(f) ShearletX: CP-L1 ↑

Figure 25: Effect of using a different transformation function in place of the proposed quantile transformation function (QTF) for the PixelMask (Fong & Vedaldi, 2017) (first row), WaveletX (Kolek et al., 2023) (second row) and ShearletX (Kolek et al., 2023) (third row). We took the square root before the min-max scaling to account for the skewness typically observed in the dataset used (see section D). The solid line represents the average of the interquartile mean (IQM) performance across 500 randomly selected validation ImageNet (Deng et al., 2009) samples, while the shaded area denotes the standard errors respectively. ↑ indicates that higher is better, and ↓ that lower is better. We employ a pretrained ResNet18 (He et al., 2016) model as a classifier.

F.12 Effect of varying λ_1 for PixelMask



(a) StartGrad (ours): CP-Pixel ↑

(b) StartGrad (ours): CP-L1 ↑

(c) Uniform: CP-Pixel ↑

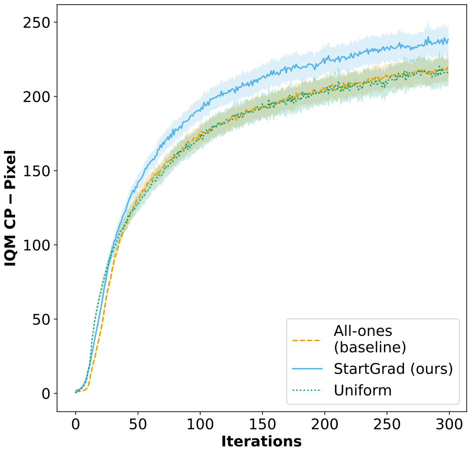
(d) Uniform: CP-L1 ↑

(e) All-ones (baseline): CP-Pixel ↑

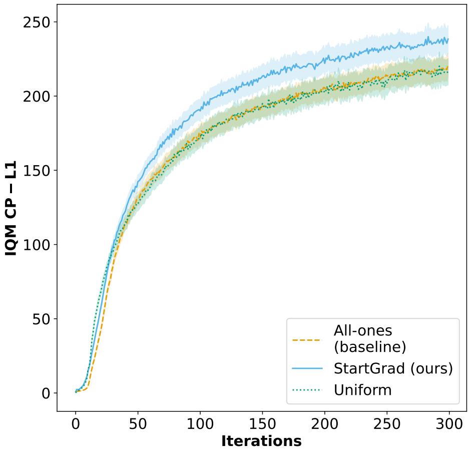
(f) All-ones (baseline): CP-L1 ↑

Figure 26: Effect of varying the λ_1 hyperparameter for the PixelMask (Fong & Vedaldi, 2017) model initialized with StartGrad (first row), uniform (second row) and all-ones (third row) on the interquartile mean (IQM) performance across 250 randomly selected validation ImageNet (Deng et al., 2009) samples for the CP-Pixel score (first column) and the CP-L1 score (second column). For each metric, ↑ indicates that higher is better, and ↓ that lower is better. We employ a pretrained ResNet18 (He et al., 2016) model as a classifier. Note that the CP-Pixel score is identical to the CP-L1 score as PixelMask (Fong & Vedaldi, 2017) framework does not apply the mask to a latent representation. Increasing λ_1 improves the CP-L1 and CP-Pixel metric across all initialization methods.

F.13 PIXELMASK PERFORMANCE $\lambda_1 = 10$



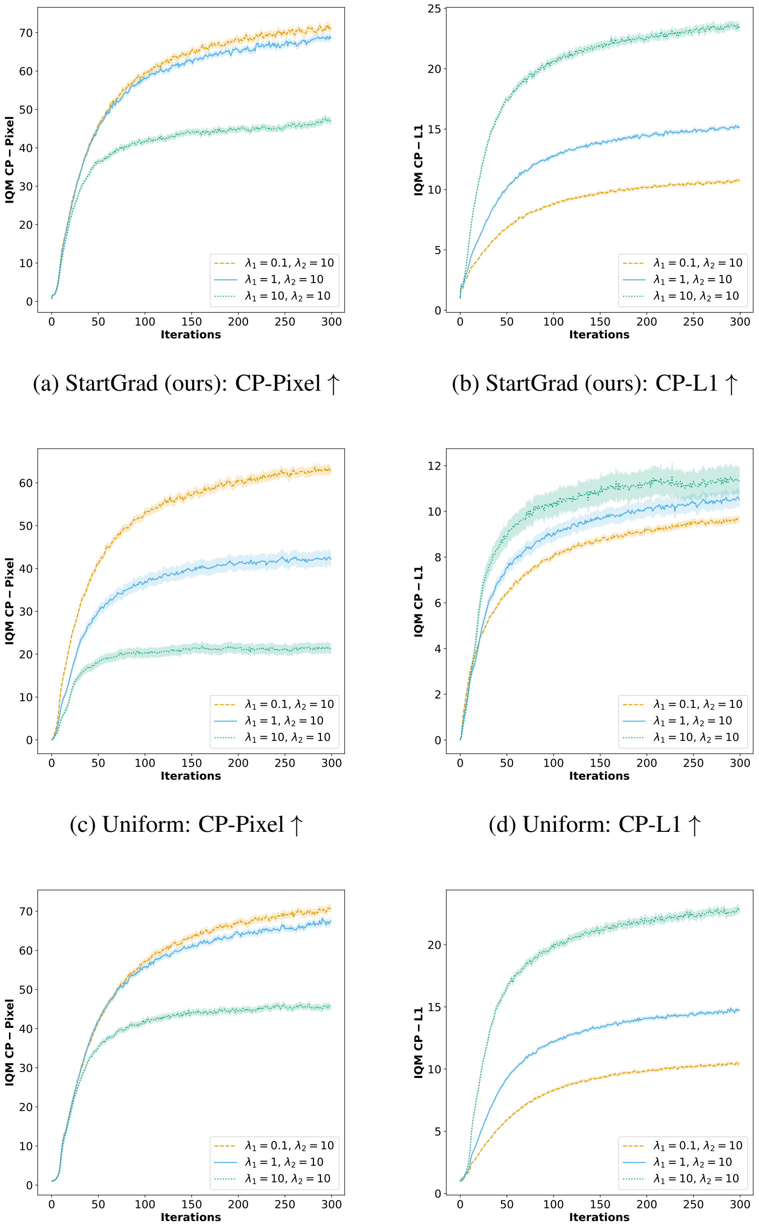
(a) PixelMask: CP-Pixel ↑



(b) PixelMask: CP-L1 ↑

Figure 27: Comparison of StartGrad initialization (ours) with standard baseline initialization schemes for the PixelMask (Fong & Vedaldi, 2017) model for $\lambda_1 = 10$. The solid line represents the average of the interquartile mean (IQM) CP-Pixel (a) and CP-L1 (b) performance across 250 randomly selected validation ImageNet (Deng et al., 2009) samples, while the shaded area denotes the standard errors respectively. For each metric, ↑ indicates that higher is better, and ↓ that lower is better. The CP-Pixel score is identical to the CP-L1 score as the mask-based model does not apply the mask to a latent representation. We employ a pretrained ResNet18 model (He et al., 2016) model as a classifier. We see that compared to all-ones and uniform, StartGrad leads to a performance boost.

F.14 Effect of varying λ_1 for WaveletX



(a) StartGrad (ours): CP-Pixel ↑

(b) StartGrad (ours): CP-L1 ↑

(c) Uniform: CP-Pixel ↑

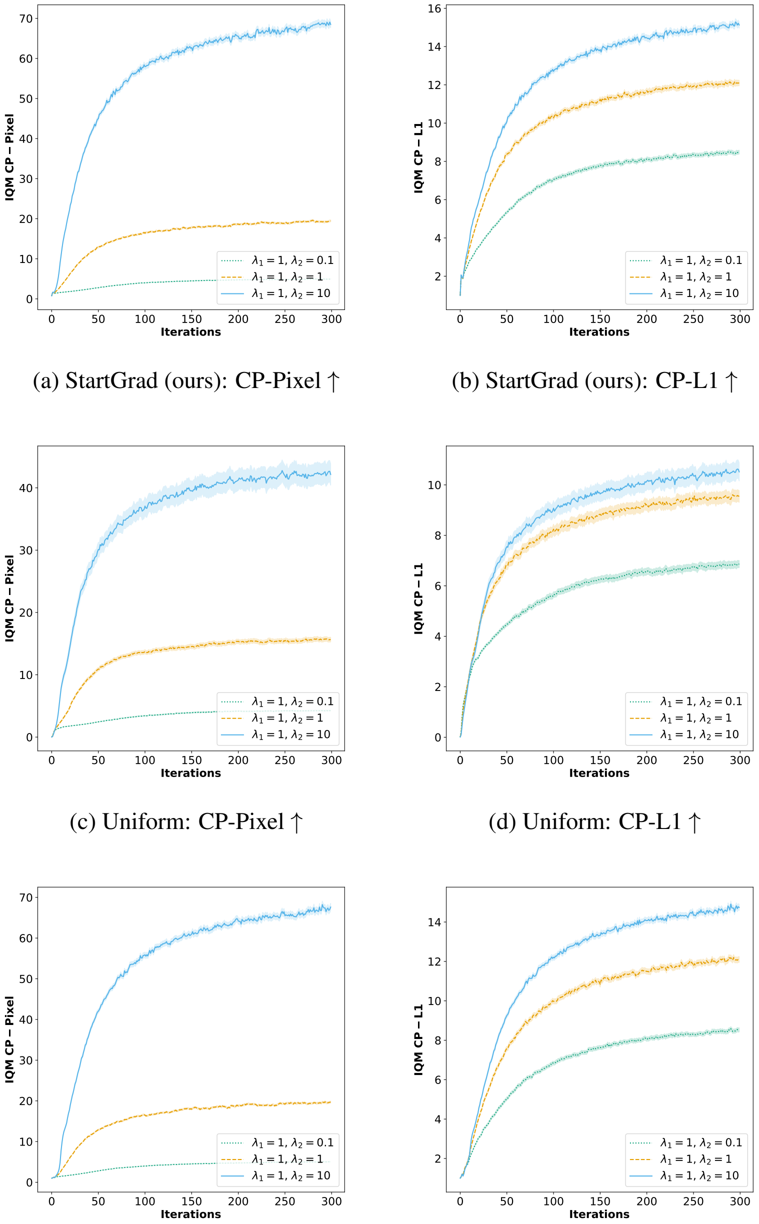
(d) Uniform: CP-L1 ↑

(e) All-ones (baseline): CP-Pixel ↑

(f) All-ones (baseline): CP-L1 ↑

Figure 28: Effect of varying the λ_1 hyperparameter for the WaveletX (Kolek et al., 2023) model initialized with StartGrad (first row), uniform (second row) and all-ones (third row) on the interquartile mean (IQM) performance across 500 randomly selected validation ImageNet (Deng et al., 2009) samples for the CP-Pixel score (first column) and the CP-L1 score (second column). For each metric, ↑ indicates that higher is better, and ↓ that lower is better. We employ a pretrained ResNet18 (He et al., 2016) model as a classifier. Increasing λ_1 improves the CP-L1 metric across all initialization methods, but decreases the CP-Pixel metric. The hyperparameter choice of $\lambda_1 = 1$ and $\lambda_2 = 10$ therefore represents the middle ground with respect to CP performance.

F.15 Effect of varying λ_2 for WaveletX



(a) StartGrad (ours): CP-Pixel ↑

(b) StartGrad (ours): CP-L1 ↑

(c) Uniform: CP-Pixel ↑

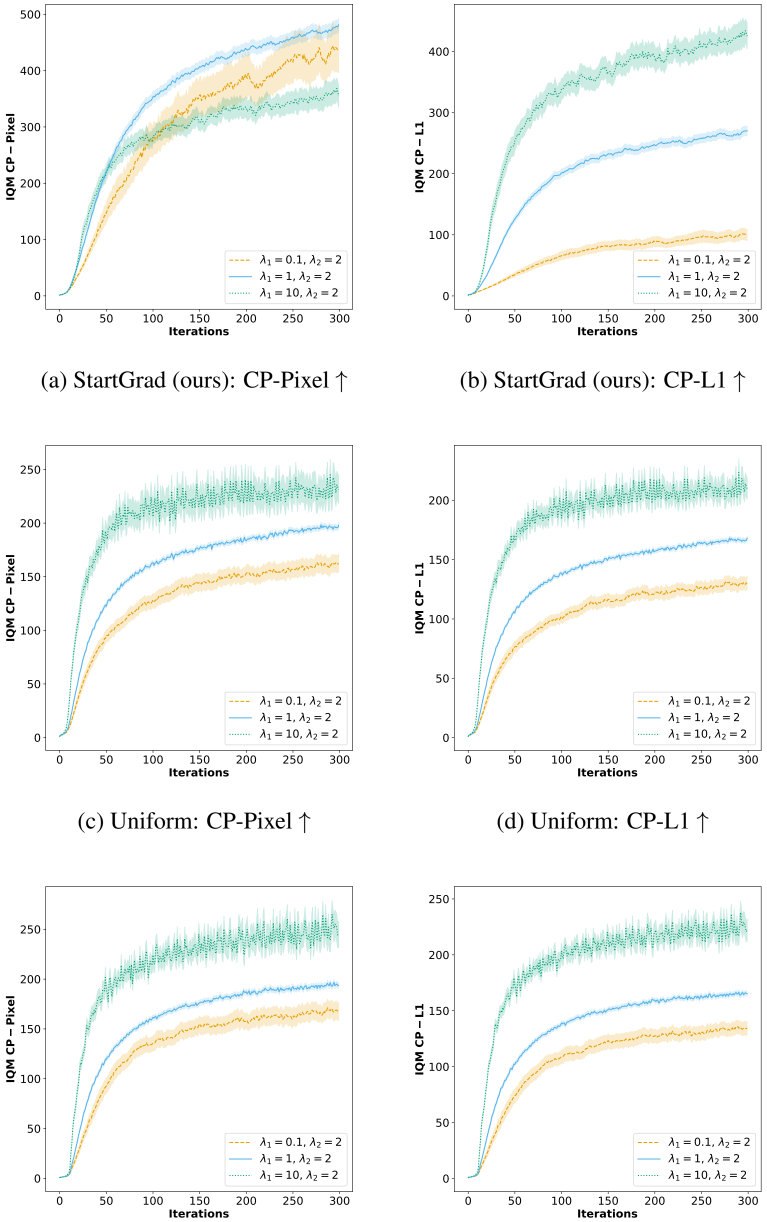
(d) Uniform: CP-L1 ↑

(e) All-ones: CP-Pixel ↑

(f) All-ones: CP-L1 ↑

Figure 29: Effect of varying the λ_2 hyperparameter for the WaveletX (Kolek et al., 2023) model initialized with StartGrad (first row), uniform (second row) and all-ones (third row) on the interquartile mean (IQM) performance across 500 randomly selected validation ImageNet (Deng et al., 2009) samples for the CP-Pixel score (first column) and the CP-L1 score (second column). For each metric, ↑ indicates that higher is better, and ↓ that lower is better. We employ a pretrained ResNet18 (He et al., 2016) model as a classifier. Increasing the λ_2 improves the scores across all initialization methods. The hyperparameter choice of $\lambda_1 = 1$ and $\lambda_2 = 10$ leads to the best overall performance across all initialization methods.

F.16 Effect of varying λ_1 for ShearletX



(a) StartGrad (ours): CP-Pixel ↑

(b) StartGrad (ours): CP-L1 ↑

(c) Uniform: CP-Pixel ↑

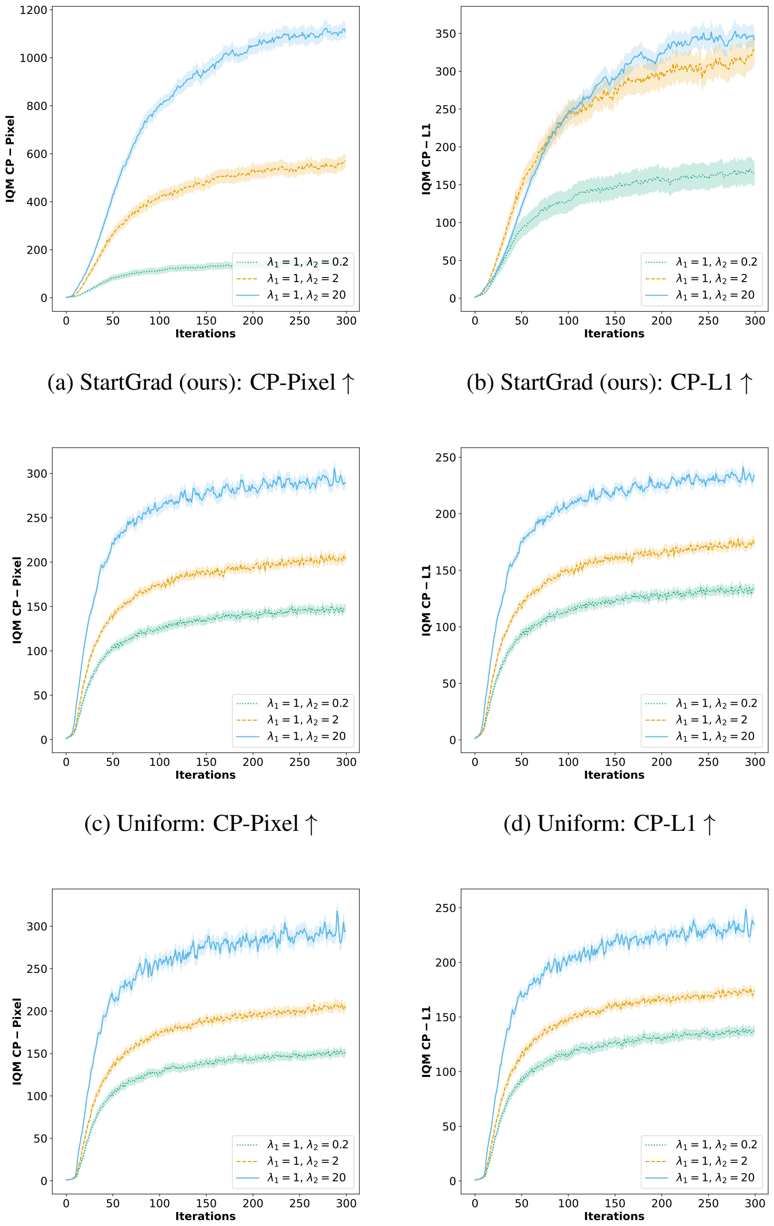
(d) Uniform: CP-L1 ↑

(e) All-ones: CP-Pixel ↑

(f) All-ones: CP-L1 ↑

Figure 30: Effect of varying the λ_1 hyperparameter for the ShearletX model initialized with StartGrad (first row), uniform (second row) and all-ones (third row) on the interquartile mean (IQM) performance for the CP-Pixel score (first column) and the CP-L1 score (second column). For each metric, ↑ indicates that higher is better, and ↓ that lower is better. We employ a pretrained ResNet18 (He et al., 2016) model as a classifier. Due to time constraints, 500 samples were used for the $\lambda_1 = 1$ and $\lambda_2 = 2$, whereas for the other hyperparameter we used a sample size of 50 explaining the larger variation. With the exception of the CP-Pixel score for the StartGrad initialization, increasing the λ_1 increases both the CP-L1 and CP-Pixel scores across all initialization methods for both metrics.

F.17 Effect of varying λ_2 for ShearletX



(a) StartGrad (ours): CP-Pixel ↑

(b) StartGrad (ours): CP-L1 ↑

(c) Uniform: CP-Pixel ↑

(d) Uniform: CP-L1 ↑

(e) All-ones: CP-Pixel ↑

(f) All-ones: CP-L1 ↑

Figure 31: Effect of varying the λ_2 hyperparameter for the ShearletX (Kolek et al., 2023) model initialized with StartGrad (first row), uniform (second row) and all-ones (third row) on the interquartile mean (IQM) performance across 500 randomly selected validation ImageNet (Deng et al., 2009) samples for the CP-Pixel score (first column) and the CP-L1 score (second column). For each metric, ↑ indicates that higher is better, and ↓ that lower is better. We employ a pretrained ResNet18 (He et al., 2016) model as a classifier. Increasing the λ_2 increases both the CP-L1 and CP-Pixel scores across all initialization methods for both metrics.

F.18 VISUALIZATIONS PIXELMASK

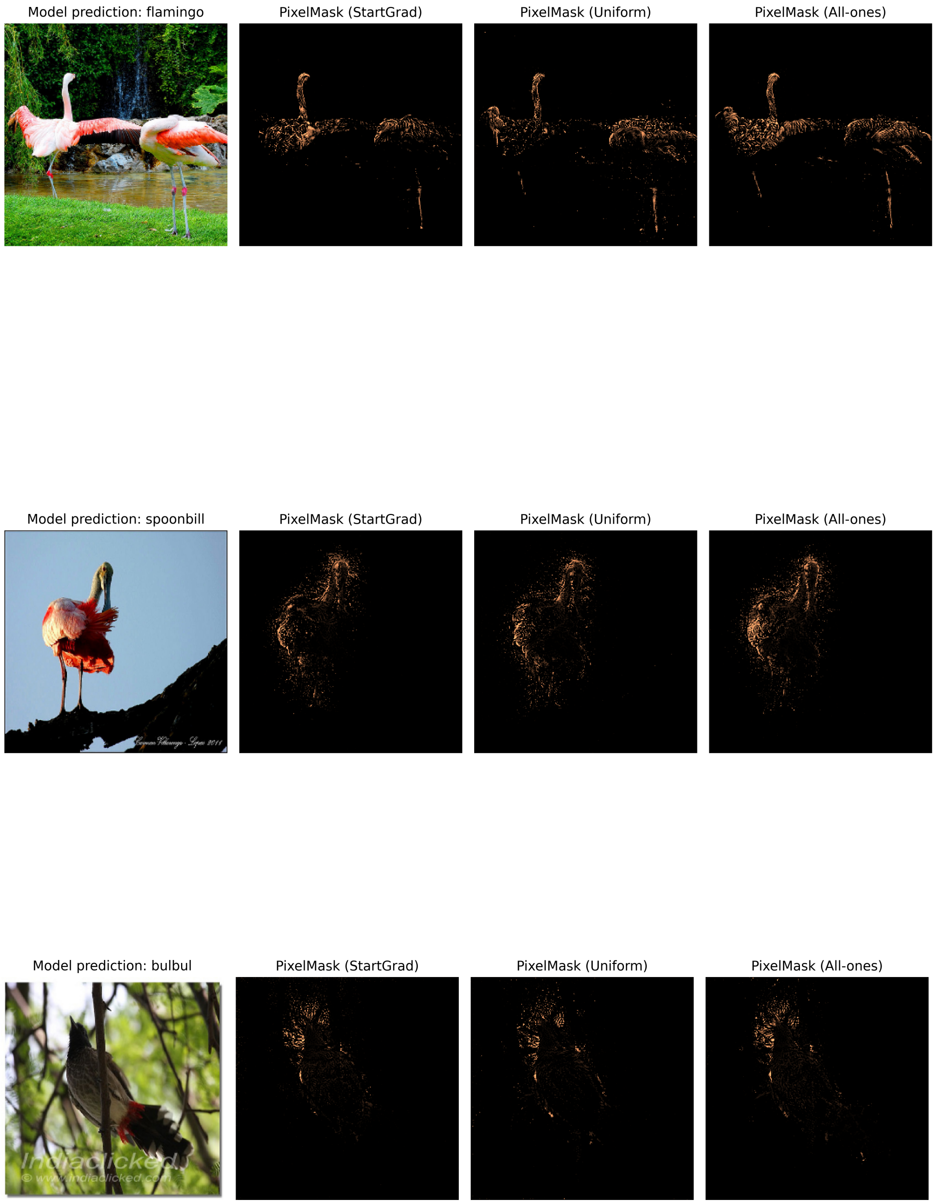


Figure 32: Illustrative visual examples of PixelMask (Fong & Vedaldi, 2017) applied to randomly selected validation ImageNet (Deng et al., 2009) samples using a pretrained ResNet18 (He et al., 2016) classifier. The optimization was run for 25 iteration steps to emphasize early visual differences between initialization methods.

F.19 Visualizations WaveletX

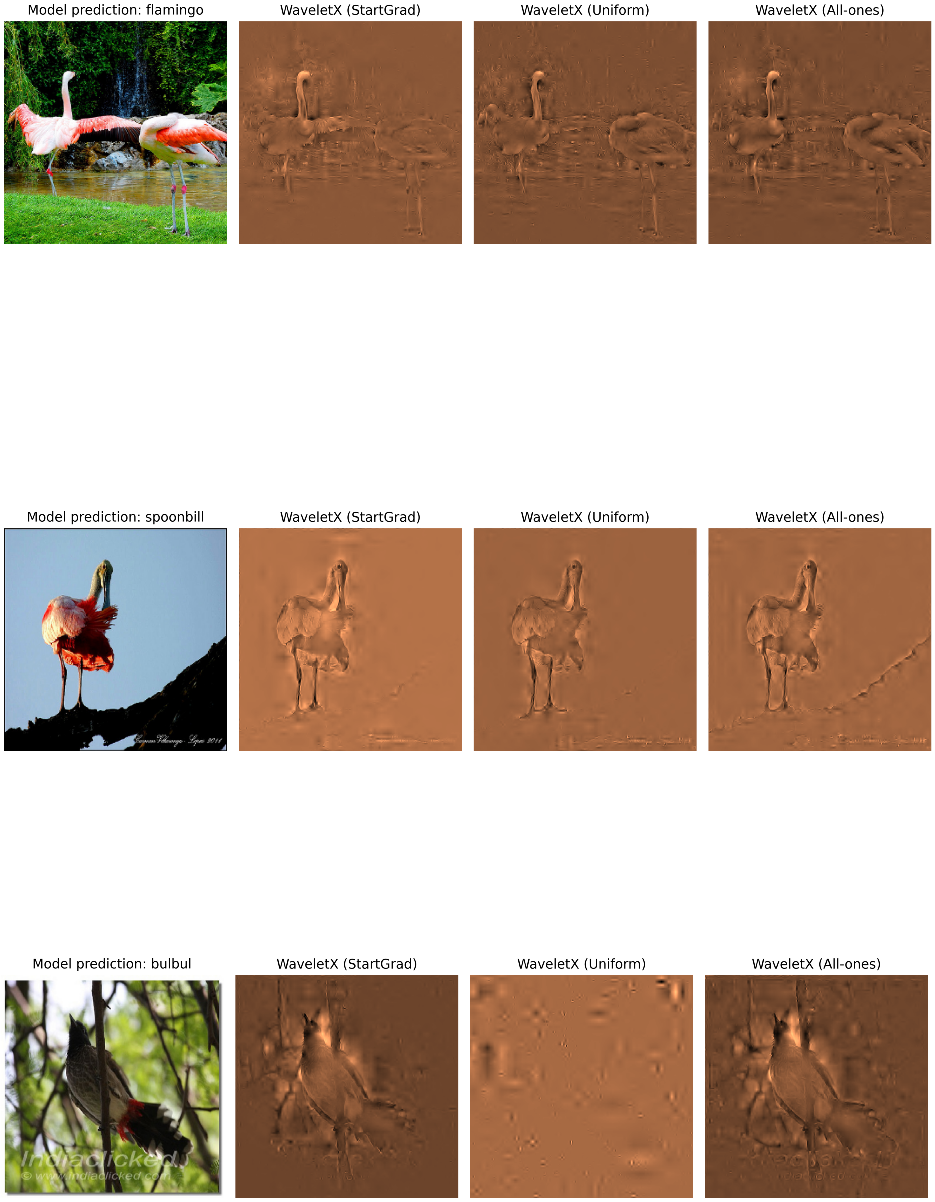


Figure 33: Illustrative visual examples of WaveletX (Kolek et al., 2023) applied to randomly selected validation ImageNet (Deng et al., 2009) samples using a pretrained ResNet18 (He et al., 2016) classifier. The optimization was run for 25 iteration steps to emphasize early visual differences between initialization methods.

F.20 VISUALIZATIONS SHEARLETX

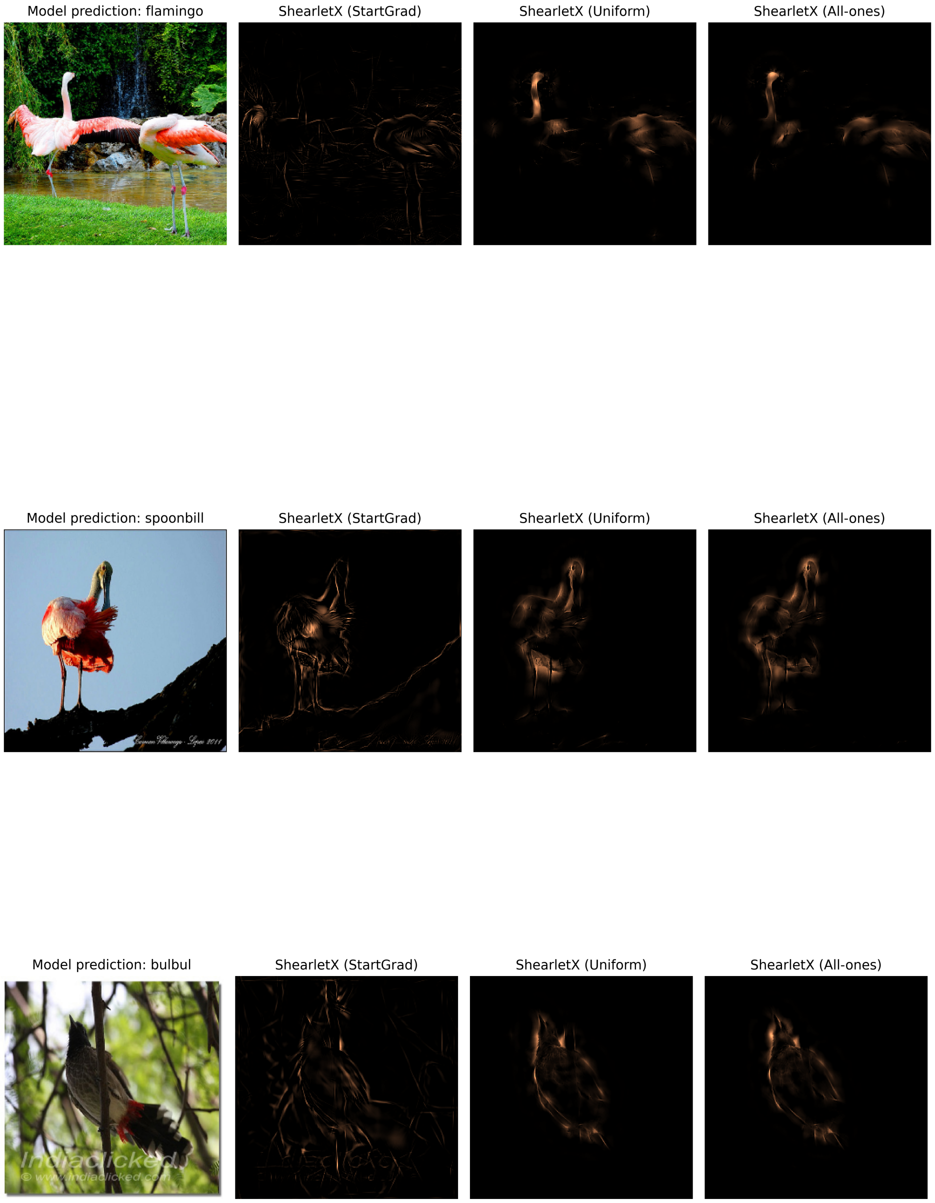
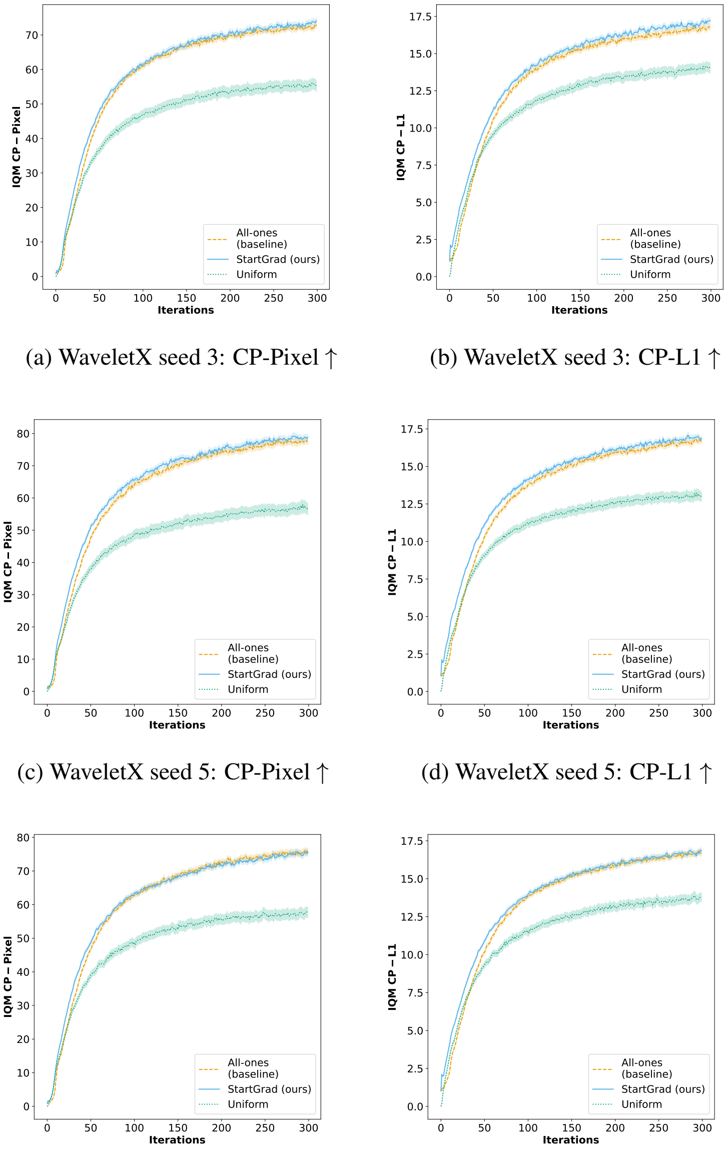


Figure 34: Illustrative visual examples of ShearletX (Kolek et al., 2023) applied to randomly selected validation ImageNet (Deng et al., 2009) samples using a pretrained ResNet18 (He et al., 2016) classifier. The optimization was run for 25 iteration steps to emphasize early visual differences between initialization methods. The results indicate that initializing with StartGrad preserves edges better, which are key for ShearletX's performance, compared to alternative initialization methods. This advantage in edge preservation likely explains StartGrad's superior performance by guiding the optimization to focus on critical features, such as edges, for improved mask-based explanations.

F.21 Random seed stability

F.21.1 Vision



(a) WaveletX seed 3: CP-Pixel ↑

(b) WaveletX seed 3: CP-L1 ↑

(c) WaveletX seed 5: CP-Pixel ↑

(d) WaveletX seed 5: CP-L1 ↑

(e) WaveletX seed 7: CP-Pixel ↑

(f) WaveletX seed 7: CP-L1 ↑

Figure 35: Comparison of StartGrad initialization (ours) with standard baseline initialization schemes for the WaveletX (Kolek et al., 2023) explanation models across different seeds. Baseline refers to the originally used initialization scheme for the respective mask explanation method. The solid line represents the average of the interquartile mean (IQM) performance across 500 randomly selected validation ImageNet (Deng et al., 2009) samples, while the shaded area denotes the standard errors respectively. For each metric, ↑ indicates that higher is better, and ↓ that lower is better. We employ a pretrained ResNet18 (He et al., 2016) model as a classifier. The initialization methods show stable behavior across random seeds: StartGrad consistently outperforms uniform initialization and performs better (seeds 3, 5) or on par (for later iteration steps seed 7) with all-ones initialization.

F.21.2 Time Series

Table 10: Average performance of the ExtremalMask method (Enguehard, 2023) (preservation game objective) across iteration steps on the state dataset using StartGrad, all-ones, and uniform initializations. Reported values are the mean and standard deviation of nine runs of the experiments using different random seeds. Each experiment averages results over five folds as stated in the main paper. Metrics are denoted with ↑ (higher is better) or ↓ (lower is better). Baseline refers to the original initialization scheme for the method. Results more than one standard deviation better than the second-best are highlighted in bold.

		Iteration steps			
Metric	Initialization	50	100	300	500
AUP ↑	Uniform (baseline)	0.723 ± 0.024	0.805 ± 0.025	0.855 ± 0.030	0.857 ± 0.030
	All-ones	**0.875 ± 0.023**	**0.845 ± 0.028**	0.862 ± 0.030	0.862 ± 0.030
	StartGrad (ours)	0.834 ± 0.031	**0.863 ± 0.031**	0.863 ± 0.031	0.863 ± 0.031
AUR ↑	Uniform (baseline)	0.692 ± 0.009	0.727 ± 0.008	0.738 ± 0.008	0.739 ± 0.007
	All-ones	0.688 ± 0.010	**0.755 ± 0.008**	0.744 ± 0.008	0.744 ± 0.008
	StartGrad (ours)	**0.795 ± 0.006**	**0.751 ± 0.008**	0.746 ± 0.008	0.746 ± 0.008
I [10^5] ↑	Uniform (baseline)	1.035 ± 0.037	2.863 ± 0.189	3.356 ± 0.120	3.376 ± 0.117
	All-ones	2.474 ± 0.098	3.269 ± 0.112	3.409 ± 0.117	3.418 ± 0.117
	StartGrad (ours)	**3.353 ± 0.163**	**3.405 ± 0.134**	3.421 ± 0.121	3.423 ± 0.121
E [10^4] ↓	Uniform (baseline)	1.866 ± 0.088	0.941 ± 0.084	0.777 ± 0.063	0.770 ± 0.063
	All-ones	1.561 ± 0.058	1.096 ± 0.059	0.773 ± 0.060	0.769 ± 0.060
	StartGrad (ours)	**1.286 ± 0.070**	**0.820 ± 0.067**	0.767 ± 0.060	0.766 ± 0.060

Table 11: Average performance of the ExtremalMask method (Enguehard, 2023) (deletion game objective) across iteration steps on the state dataset using StartGrad, all-ones, and uniform initializations. Reported values are the mean and standard deviation of nine runs of the experiments using different random seeds. Each experiment averages results over five folds as stated in the main paper. Metrics are denoted with ↑ (higher is better) or ↓ (lower is better). Baseline refers to the original initialization scheme for the method. Results more than one standard deviation better than the second-best are highlighted in bold.

		Iteration steps			
Metric	Initialization	50	100	300	500
AUP ↑	Uniform (baseline)	0.648 ± 0.051	0.721 ± 0.050	0.817 ± 0.062	0.853 ± 0.064
	All-ones	**0.818 ± 0.084**	**0.880 ± 0.083**	0.884 ± 0.072	0.886 ± 0.070
	StartGrad (ours)	**0.815 ± 0.062**	0.853 ± 0.074	0.871 ± 0.072	0.875 ± 0.071
AUR ↑	Uniform (baseline)	0.674 ± 0.010	0.712 ± 0.009	0.743 ± 0.012	0.743 ± 0.013
	All-ones	0.459 ± 0.023	0.601 ± 0.015	0.739 ± 0.011	0.741 ± 0.012
	StartGrad (ours)	**0.804 ± 0.009**	**0.751 ± 0.012**	0.740 ± 0.012	0.739 ± 0.012
I [10^5] ↑	Uniform (baseline)	1.112 ± 0.072	3.888 ± 0.067	4.217 ± 0.083	4.256 ± 0.092
	All-ones	0.664 ± 0.056	0.932 ± 0.047	4.223 ± 0.078	4.268 ± 0.084
	StartGrad (ours)	**4.245 ± 0.070**	**4.270 ± 0.083**	4.267 ± 0.084	4.266 ± 0.084
E [10^4] ↓	Uniform (baseline)	1.523 ± 0.101	0.493 ± 0.019	0.386 ± 0.016	0.338 ± 0.020
	All-ones	1.753 ± 0.047	1.838 ± 0.071	0.343 ± 0.026	0.317 ± 0.026
	StartGrad (ours)	**1.004 ± 0.033**	**0.414 ± 0.015**	**0.311 ± 0.021**	0.304 ± 0.021

Table 12: Average performance of the ExtremalMask method (Enguehard, 2023) (preservation game objective) across iteration steps on the switch-feature dataset using StartGrad, all-ones, and uniform initializations. Reported values are the mean and standard deviation of nine runs of the experiments using different random seeds. Each experiment averages results over five folds as stated in the main paper. Metrics are denoted with ↑ (higher is better) or ↓ (lower is better). Baseline refers to the original initialization scheme for the method. Results more than one standard deviation better than the second-best are highlighted in bold.

		Iteration steps			
Metric	Initialization	50	100	300	500
AUP ↑	Uniform (baseline)	0.878 ± 0.007	0.964 ± 0.003	0.984 ± 0.004	0.985 ± 0.004
	All-ones	**0.947 ± 0.002**	0.961 ± 0.006	0.984 ± 0.004	0.986 ± 0.003
	StartGrad (ours)	0.920 ± 0.008	**0.982 ± 0.003**	0.986 ± 0.003	0.986 ± 0.003
AUR ↑	Uniform (baseline)	0.724 ± 0.012	0.725 ± 0.013	0.733 ± 0.018	0.732 ± 0.017
	All-ones	0.689 ± 0.010	**0.761 ± 0.014**	0.739 ± 0.018	0.738 ± 0.018
	StartGrad (ours)	**0.799 ± 0.016**	**0.747 ± 0.019**	0.739 ± 0.019	0.739 ± 0.018
I $[10^5]$ ↑	Uniform (baseline)	1.032 ± 0.037	2.039 ± 0.149	2.661 ± 0.194	2.668 ± 0.198
	All-ones	1.978 ± 0.059	**2.597 ± 0.162**	2.678 ± 0.198	2.674 ± 0.200
	StartGrad (ours)	**2.617 ± 0.183**	**2.668 ± 0.193**	2.675 ± 0.200	2.674 ± 0.198
E $[10^4]$ ↓	Uniform (baseline)	2.266 ± 0.098	1.487 ± 0.080	1.126 ± 0.109	1.162 ± 0.114
	All-ones	1.933 ± 0.034	1.475 ± 0.075	1.137 ± 0.110	1.137 ± 0.109
	StartGrad (ours)	**1.604 ± 0.104**	**1.232 ± 0.099**	1.138 ± 0.107	1.136 ± 0.107

Table 13: Average performance of the ExtremalMask method (Enguehard, 2023) (deletion game objective) across iteration steps on the switch-feature dataset using StartGrad, all-ones, and uniform initializations. Reported values are the mean and standard deviation of nine runs of the experiments using different random seeds. Each experiment averages results over five folds as stated in the main paper. Metrics are denoted with ↑ (higher is better) or ↓ (lower is better). Baseline refers to the original initialization scheme for the method. Results more than one standard deviation better than the second-best are highlighted in bold.

		Iteration steps			
Metric	Initialization	50	100	300	500
AUP ↑	Uniform (baseline)	0.835 ± 0.039	0.894 ± 0.050	0.930 ± 0.060	0.935 ± 0.058
	All-ones	0.764 ± 0.075	0.751 ± 0.074	0.804 ± 0.086	0.805 ± 0.093
	StartGrad (ours)	**0.880 ± 0.024**	**0.926 ± 0.031**	0.949 ± 0.027	0.954 ± 0.026
AUR ↑	Uniform (baseline)	0.602 ± 0.029	**0.598 ± 0.029**	**0.607 ± 0.032**	**0.599 ± 0.032**
	All-ones	0.638 ± 0.054	0.527 ± 0.025	0.554 ± 0.039	0.547 ± 0.043
	StartGrad (ours)	**0.670 ± 0.020**	**0.607 ± 0.019**	**0.598 ± 0.012**	**0.595 ± 0.011**
I $[10^5]$ ↑	Uniform (baseline)	0.763 ± 0.053	1.620 ± 0.189	**2.139 ± 0.211**	**2.107 ± 0.188**
	All-ones	0.830 ± 0.099	0.553 ± 0.028	1.354 ± 0.180	1.570 ± 0.277
	StartGrad (ours)	**1.740 ± 0.094**	**1.987 ± 0.096**	**2.100 ± 0.055**	**2.098 ± 0.044**
E $[10^4]$ ↓	Uniform (baseline)	2.500 ± 0.084	1.685 ± 0.097	**1.388 ± 0.036**	**1.365 ± 0.031**
	All-ones	2.339 ± 0.073	2.860 ± 0.101	1.889 ± 0.079	1.649 ± 0.099
	StartGrad (ours)	**2.404 ± 0.028**	**1.613 ± 0.051**	**1.354 ± 0.036**	**1.336 ± 0.036**