Group-Sparse Manifold-Aware Integrated Gradients for Multimodal Transformers on EHR Trajectories

Ali Amirahmadi ALI.AMIRAHMADI@HH.SE
Halmstad University, Sweden

Farzaneh Etminani FARZANEH.ETMINANI@HH.SE
Halmstad University, Sweden
Region Halland, Sweden

Mattias Ohlsson MATTIAS.OHLSSON@CEC.LU.SE
Halmstad University, Sweden
Lund University, Sweden

Abstract

Integrated Gradients (IG) is a popular method for explaining clinical deep models—including widely used multimodal, pretrained Transformers—but its utility on EHR code sequences is hampered by (i) the lack of principled baselines for sequence of discrete tokens and (ii) dense, hard-to-interpret generated attributions. To address both, first, we introduce a manifold-aware baseline: the mean input embedding (computed on the validation set), which keeps IG's interpolated points close to typical sequences in embedding space. Second, we introduce GS-IG, which preserves the straight path geometry but re-parameterizes the schedule $\alpha(t) = t^\theta$ and selects θ per input by minimizing a token-level $\ell_{2,1}$ (group-sparsity) objective, producing concise, practitioner-friendly explanations. On MIMIC-IV (incident heart failure) and MDC (early mortality), the manifold-aware baseline improves faithfulness (higher Comprehensiveness, lower Sufficiency), and GS-IG reduces token-level $\ell_{2,1}$ by 9–18% with negligible change in those metrics on the manifold-aware baseline. The method is lightweight and yields faithful, sparse, and actionable explanations.

Keywords: Integrated Gradients, Explainability, Multimodal Transformers, Group Sparsity, Manifold-aware, Electronic Health Records (EHR), Patient trajectories

Data and Code Availability This research has been conducted using the MIMIC-IV (v2.2) and the Malmö Diet and Cancer (MDC) Cohort data. MIMIC-IV (v2.2) is publicly available at https://physionet.org/content/mimiciv/2.2/. The MDC data used in this study are not publicly available due to restrictions imposed by the Malmö Population-Based Cohorts Joint Database; data may be requested with appropriate approvals from the database (https://www.malmo-kohorter.lu.se/malmo-cohorts). The code is available at https://github.com/ali-amirahmadii/Group-Sparse-IG;

Institutional Review Board (IRB) Research on de-identified data from MIMIC-IV is exempt from IRB review under HIPAA. Use of the Malmö Diet and Cancer (MDC) Cohort was approved by the Swedish Ethical Review Authority (Dnr 2023-00503-01).

1. Introduction

Deep neural networks—particularly Transformer architectures in large foundation models—have achieved state-of-the-art, and in many cases superhuman, performance across a broad range of vision, language, and multimodal benchmarks Bommasani et al. (2021); Guo et al. (2023); Wornow et al. (2023); Amirahmadi et al. (2025b). Yet in risk-sensitive settings the value of a model hinges not only on how often it is correct, but also on *why* it issues a given decision. Post-hoc attribution methods therefore play a crucial role in turning black-box predictions into *actionable insight*. Among these methods, Integrated Gradients(IG) (Sundararajan et al., 2017) is especially popular because it addresses the gradient saturation challenge through line integration and satisfies key axioms (Sensitivity, Implementation Invariance, Linearity, and completeness). See 2.1 and A.11.

© 2025 A. Amirahmadi, F. Etminani & M. Ohlsson.

However, two important design choices restrict IG's effectiveness on non-image architectures that operate in embedding space. The baseline and the integration path from the baseline to the input. Straight-line IG assumes access to an input that represents "absence of evidence". While a black canvas or zero vector suffices for pixels, categorical sequences—sentences or streams of medical codes—lack an obvious null token. Recent NLP work resorts to special embeddings such as `<MASK>` or explains one token at a time (Sequential IG Enguehard (2023)) to keep interpolates meaningful, but these synthetic heuristics can push paths off the data manifold and leave the global baseline ill-defined. Path choice also matters. Straight lines in embedding space may traverse low-density regions, amplifying gradient noise and reducing faithfulness. Moreover, downstream users typically prefer a concise list of decisive factors over a dense saliency map.

We revisit IG through the lens of data-manifold alignment and group sparsity. Our central insight is that, for models operating in an embedding space with natural groups (e.g., words, diagnosis codes), the baseline should consider that structure.

Manifold-aware baseline. We take the mean embedding of validation-set patient trajectories—averaging token vectors while preserving positions—and use these vectors as a principled 'null' point. By manifold-aware, we mean near the learned embedding distribution, i.e., close to the high-density regions formed by embeddings the model produces on real validation data, rather than synthetic, low-density special tokens.

Group-Sparse prior path. Inspired by adaptive-path methods, we keep the straight IG path and only reparameterize its speed via the schedule $\alpha(t) = t^\theta$. For each input, we select the θ that encourages token-level sparsity. Entire embeddings (tokens) therefore rise or fall together, yielding token-level rather than dimension-level attributions.

Contributions

- **Manifold-aware baseline for discrete sequences.** We propose a position-wise empirical mean of token embeddings to anchor IG near the model's data manifold.
- **Path-optimized, group-sparse IG.** We retain straight-path geometry but re-parameterize the schedule $\alpha(t) = t^\theta$ and select θ per input by minimizing a token-level $\ell_{2,1}$ objective, yielding concise, group-sparse attributions while preserving IG's axioms.
- **Empirical gains on two EHR cohorts.** On a multimodal Transformer for early mortality (MDC) and incident heart failure (MIMIC-IV), the manifold-aware baseline improves faithfulness (Comprehensiveness↑, Sufficiency↓); adding GS-IG matches vanilla IG on fidelity while reducing token-level group sparsity by $\approx$ –18% on the manifold-aware baseline.

By disentangling baseline selection, path scheduling, and sparsity priors, our approach turns IG into an explainer for modern multimodal Transformers, delivering faithful, concise, and actionable attributions wherever inputs are represented as learned embeddings.

2. Methods

We first review the notation and the classical Integrated-Gradients (IG) framework, then introduce a manifold-aware baseline that is suited to categorical sequences processed by a transformer, and finally present our Group Sparse path-optimised IG (**GS-IG**) algorithm, which couples an adaptive integration schedule on the straight path with a group-lasso prior to obtain concise, token-level attributions.

2.1. Preliminaries

2.1.1. Notation and definitions

Let $F : \mathbb{R}^d \to \mathbb{R}$ denote a trained neural network whose scalar output we wish to explain; when the model has a vector output, we treat each component separately. A single input example is the multimodal tuple $x = \left(x^{(1)}, \ldots, x^{(M)}\right)$, where M is the number of modalities under consideration. For a sequence modality—such as a sentence or a stream of diagnosis codes—the j-th discrete token is mapped by the embedding layer to $e_j \in \mathbb{R}^{d_e}$. Also, continuous modalities (e.g. age, laboratory values) are left in their native Euclidean spaces.

Integrated-Gradients methods require a baseline input x' that represents the absence of informative features. Together with x it defines a differentiable path $\gamma : [0, 1] \to \mathbb{R}^{d_e}$ satisfying $\gamma(0) = x'$ and $\gamma(1) = x$. Throughout, we assume that the d input coordinates can be partitioned into a collection of disjoint *groups*, $\mathcal{G} = \{G_1, \ldots, G_{|\mathcal{G}|}\}$. All d_e coordinates of a single token embedding belong to the same group, which will allow us to impose sparsity at the

token—rather than dimension—level in subsequent sections.

2.1.2. Integrated Gradients

IG (Sundararajan et al., 2017) explains a prediction by integrating the input–output sensitivity of the model along a continuous path, that connects the baseline x' to the observation x. The attribution assigned to the i^{th} input coordinate is

$$\mathrm{IG}_i(x; x', \gamma) \;=\; \int_0^1 \frac{\partial F\big(\gamma(t)\big)}{\partial \gamma_i(t)} \, \frac{\partial \gamma_i(t)}{\partial t} \, dt, \qquad (1)$$

which reduces to the familiar straight-line formulation when $\gamma(t) = x' + t\,(x - x')$. Equation (1) satisfies the completeness axiom, $\sum_i \mathrm{IG}_i = F(x) - F(x')$, ensuring that the total attribution equals the change in the model output between the baseline and the input.

In practice, the integral is approximated with a Riemann sum over K samples t_k:

$$\begin{aligned} \widehat{\mathrm{IG}}_i(x; x', \gamma) &= \sum_{k=1}^{K} g_{k,i}\, \Delta\gamma_{k,i}, \\ g_k &= \nabla_x F\big(\gamma(t_k)\big), \\ \Delta\gamma_{k,i} &= \gamma_i(t_k) - \gamma_i(t_{k-1}). \end{aligned} \qquad (2)$$

Here, $\gamma(t_k)$ denotes the k-th intermediate input (a point in input space) along the path from the baseline x' to the input x, and $\gamma_i(t_k)$ is the i-th feature of this point.

2.2. Manifold-Aware Baseline

IG explanations are known to be highly sensitive to the choice of baseline. Whereas a zero image or silence frame can serve as a natural "null" input for vision or audio, defining an absence-of-evidence point for a sequence of categorical tokens is less obvious. Prior work often uses sequences of special symbols such as `<PAD>` or `<MASK>`; however, these are synthetic and can lie outside the empirical data manifold, and may yield noisy or even misleading interpolations (Enguehard, 2023; Kapishnikov et al., 2021). Figures 1 illustrate that commonly used baselines lie far from the model's learned embedding manifold for medical trajectories from the MDC and MIMIC-IV datasets (See 3.1).

To alleviate this problem we replace discrete dummy tokens with a manifold-aligned baseline constructed directly from the validation data. For modality m, we pass each validation sequence through the transformer and collect the embeddings $h_{n,j}^{(m)} \in \mathbb{R}^{d_e}$ produced for token j in sample n. Those token states are concatenated in their original order to form a single *sequence-level embedding*,

$$E_n^{(m)} = \big[h_{n,1}^{(m)} \,\|\, h_{n,2}^{(m)} \,\|\, \cdots \,\|\, h_{n,L_m}^{(m)}\big] \in \mathbb{R}^{L_m d_e},$$

where L_m is the length of the sequence and d_e is the embedding dimension. Computing the centroid in embedding space across the validation set, the emperical mean,

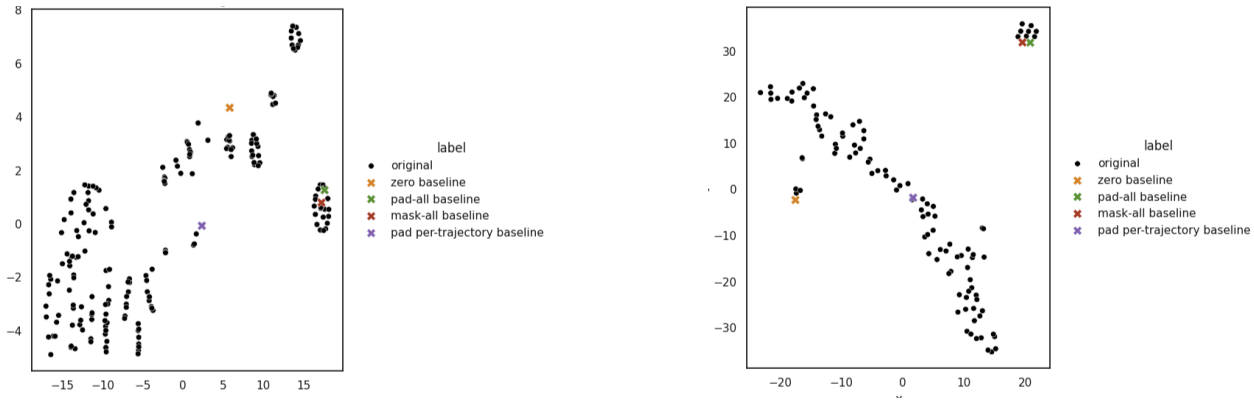
$$\mu^{(m)} \;=\; \frac{1}{N} \sum_{n=1}^{N} E_n^{(m)}, \qquad (3)$$

yields a single baseline that (i) lies near the support of the learned embedding distribution, (ii) preserves positional structure through concatenation, and (iii) is free of heuristics tied to any particular vocabulary item. The vector $\mu^{(m)}$ is reused for every test sequence belonging to modality m, providing a stable and semantically grounded origin for the IG path integral. For the numerical features we adopt the conventional baseline $x'^{(m)} = \mathrm{mean}_n(x_n^{(m)})$. Then IGs are computed per modality using the corresponding baseline and then concatenated to yield a unified attribution vector.

Why it helps: IG attributes the change in prediction along a path from a baseline x' to the input x. When x' is a special token (e.g., `<PAD>`, the straight-line path traverses low-density regions of the learned embedding distribution, which can amplify gradient noise. Selecting x' as the mean of the embeddings (Eq. (3)) keeps the interpolates near higher-density neighborhoods and yields less noisy accumulation. See Appendix A.10 for an empirical analysis showing that the mean-embedding baseline lies in higher-support regions of the sequence-embedding space and improves IG path quality.

2.3. Group-Sparse IG (GS-IG)

Conventional gradient-based explanation methods, including vanilla IG, usually return dense saliency scores that are difficult to read and even harder to act upon (Heo et al., 2019; Ghorbani et al., 2019; Zhang and Farnia, 2023). We propose Group-Sparse IG (GS-IG), which produces compact, token-level explanations by re-parameterizing the IG path's speed/schedule and selecting its schedule to encourage group sparsity across each token's embedding coordinates. In contrast to MoreauGrad—which



(a) MDC dataset (b) MIMIC-IV dataset

Figure 1: t-SNE visualization of input embeddings and baseline reference points (see Section 3.3.1). Some traditional baselines may align with a small sub-cluster but do not represent the majority of samples in the model's embedding space.

smooths the class-score function via a Moreau envelope and can impose ℓ_1 or $\ell_{2,1}$ penalties to produce sparse or group-sparse interpretations (Zhang and Farnia, 2023)—our method remains within the IG family: we don't compute a proximal inner problem; instead, we choose the integration schedule that makes IG naturally sparse at the token level.

One-parameter path schedule. We replace the straight-line schedule $\alpha(t) = t$ with

$$\gamma_\theta(t) \;=\; x' + \alpha_\theta(t)\,(x - x'), \alpha_\theta(t) = t^\theta, \quad \theta \in [0.1, 5.0], \tag{4}$$

where the exponent θ smoothly interpolates between baseline-biased ($\theta > 1$) and input-biased ($\theta < 1$) trajectories (Fig. 2.3). We discretize $t \in [0, 1]$ at K points and form a Riemann sum in α as in Eq. (2). Unlike Kapishnikov et al. (2021), who alter the geometry of the path in latent space, Eq. (4) keeps the geometry fixed but changes its schedule; This single scalar degree of freedom is cheap to search and empirically sufficient to promote sparser explanations.

Group Sparse objective over tokens. Let $\widehat{\mathrm{IG}}_{\mathrm{emb}}(x, x'; \theta) \in \mathbb{R}^{L \times d_e}$ denote the IG attribution matrix in embedding space (Eq. (1)), where row j corresponds to token position j. We define token saliency via the group norm $s_j = \left\|\widehat{\mathrm{IG}}_{\mathrm{emb}}[j, :]\right\|_2$ and select θ by minimizing an $\ell_{2,1}$, Group Lasso Yuan and Lin (2006), penalty over all token's embedding vectors and a ℓ_1, Lasso penalty, element-wise penalty:

$$\mathcal{L}(\theta) = \lambda_{\mathrm{grp}} \sum_{j=1}^{L} m_j \left\|\widehat{\mathrm{IG}}_{\mathrm{emb}}[j, :]\right\|_2 \;+\; \lambda_1 \left\|\widehat{\mathrm{IG}}_{\mathrm{emb}}\right\|_1, \tag{5}$$

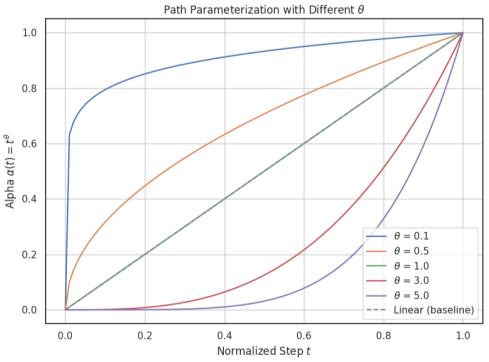


Figure 2: Effect of θ on the IG traversal schedule from baseline (α=0) to input (α=1). $\theta > 1$ lingers near the baseline (baseline-biased); $\theta < 1$ spends more steps near the input (input-biased). Equal steps in t therefore correspond to unequal steps in embedding space, changing where gradients are sampled and accumulated along the path.

where $m_j \in \{0,1\}$ masks out padding/special tokens. The first term promotes token-level sparsity (entire embeddings switch off together), aligning the explanation granularity with discrete inputs (words, medical codes). The rest of this paper we only use the first term, group sparsity objective which suits our goals best. For more details on how group Lasso helps token sparsity, please see Appendix A.1

Sample-efficient path selection. Because θ is a scalar, we choose it per case via a fast Bayesian search over $\theta \in [0.1, 5.0]$ with K=50 path samples and T=10 trials. Each trial computes IG along the hook-injected path (Sec. 2.3) and evaluates $\mathcal{L}(\theta)$; we keep $\theta^\star = \arg\min_\theta \mathcal{L}(\theta)$. This adds T IG runs per instance—lightweight relatively.

Outcome. GS-IG turns dense embedding-level attributions into a concise list of influential tokens through a tunable path schedule and a group-sparse objective, yielding faithful yet compact explanations that align with how clinicians reason about medical codes—and, more broadly, with domain experts' preference for sparse, actionable interpretations. By aligning baseline, path, and sparsity with the natural group structure of embedding spaces, GS-IG delivers concise, token-level explanations while retaining the theoretical guarantees of Integrated Gradients. As a path-based attribution method, GS-IG satisfies the standard IG axioms—completeness, sensitivity (a/b), and implementation invariance (Sundararajan et al., 2017). Notably, GS-IG preserves completeness because, prior to any grouping or sparsification, it computes a standard path integral of the gradient between two endpoints; re-parameterizing the path's speed does not change the value of that integral.

3. Experiments

3.1. Datasets and Tasks

We evaluate on two distinct cohorts and different prediction tasks: MIMIC-IV (Johnson et al., 2020) for Heart Failure (HF) prediction using both temporal, sequential EHR codes and static tabular demographic features; and the Malmö Diet and Cancer (MDC) dataset (Berglund et al., 1993) for early-death prediction using sequential medical codes and static tabular risk factors. As MDC lacks other modalities (e.g., notes or imaging), we focus only on these two most common modalities to evaluate the generalizability of our IG-based approach. Further dataset specifications and problem formulations are in Appendix A.3 and A.4.

3.2. Evaluation Metrics

We evaluate the faithfulness of model explanations with Comprehensiveness (Comp) and Sufficiency (Suff), two widely used metrics in explainability research (DeYoung et al., 2019; Sanyal and Ren, 2021; Enguehard, 2023). Comp measures the drop in model performance when the most k% important features (as determined by an attribution method) are removed, where a greater drop indicates a more faithful explanation. Conversely, Suff evaluates how well the model retains its confidence when only the most important features are retained, with a smaller drop suggesting that the model relies primarily on these features for decision-making. Further details about Comp and Suff are provided in Appendix A.5.

3.3. Experimental Setup

We train a multi-modal Transformer encoder (MMT) to predict early death and first occurrence of Heart Failure (HF), on the two datasets described in Sec. 3.1. For robustness, we first pretrain a Transformer encoder on all training-set medical codes using TOO-BERT (Amirahmadi et al., 2025a), then fine-tune it for each target task using the training/validation splits, and report all metrics on the held-out test sets. Architectural details of the MMT are provided in the Appendix A.2.

3.3.1. Manifold-aware baseline

We compare IG under several baseline choices for both temporal categorical sequences and non-temporal (tabular) features, and we also evaluate them against two recent IG variants designed for transformer encoders and baseline selection robustness.

Given prior reports that IG often exhibits stronger faithfulness than DeepLIFT (Shrikumar et al., 2017) and GradientSHAP (Lundberg, 2017) in related settings (Sanyal and Ren, 2021; Enguehard, 2023), we focus primarily on IG with the baseline variants above.

Methods compared. (i) **Gradient×Input** (Simonyan et al., 2013; Shrikumar et al., 2016): element-wise product of inputs and their gradients. (ii) **IG (zero baseline)**: a zero tensor in embedding space for sequences and a zero vector for tabular features.

(iii) **IG (pad-all)**: replace all medical codes with `<pad>` tokens before integrating. (iv) **IG (mask-all)**: replace all medical codes with `<mask>` tokens before integrating. (v) **IG (Pad equal to each trajectory baseline)**: replace all codes with `<pad>` while *preserving* the original number of tokens per visit. (vi) **Sequential Integrated Gradients (SIG)** (Enguehard, 2023): token-level IG computed sequentially, tailored to Transformer/Large LM architectures. (vii) **Expected Gradients (EG)** (Erion et al., 2021): an IG variant designed to be less sensitive to baseline. (viii) **IG with manifold-aware baseline (in embedding space)**: our manifold-aligned baseline from Sec. 2.2.

3.3.2. Group-sparse, path-optimized IG.

We assess the effect of the proposed GS-IG (Sec. 2.3) by comparing IG with and without path optimization across all baselines. Group sparsity is measured over tokens using the sum of token-level ℓ_2 norms of embedding attributions (group $\ell_{2,1}$); For an input (x, x'), let

$$S(\theta) \;=\; \big\|\widehat{\mathrm{IG}}_{\mathrm{emb}}(x, x'; \theta)\big\|_{2,1} \;=\; \sum_{j=1}^{L} m_j \,\big\|\widehat{\mathrm{IG}}_{\mathrm{emb}}[j, :]\big\|_2$$

be the token-level group-sparsity measure (mixed $\ell_{2,1}$ norm; m_j masks padding/special tokens). We report the *relative reduction* in group sparsity achieved by the optimized schedule $\theta^\star$ (found as in Sec. 2.3) compared to the straight-line schedule (θ=1):

$$\mathrm{RR}_{2,1} \;=\; 100 \times \frac{S(\theta{=}1) - S(\theta^\star)}{S(\theta{=}1)} \;\%;$$

4. Results & Discussion

We tuned the MMT, pretrained with TOO-BERT, on the two described tasks. Comprehensive performance results by task and modality are reported in Appendix A.2.1.

4.1. Impact of the Manifold-Aware Baseline on IG

After training the MMT for each task, we evaluated explanation faithfulness (Tables 1 and Appendix Table 9) using Comprehensiveness (Comp; higher is better) and Sufficiency (Suff; lower is better). IG with proposed manifold-aware baseline consistently outperforms other variants across both modalities, and IG-based methods generally outperform Gradient×Input. This underscores the central role of a realistic, data-aligned baseline in attribution quality. Appendix A.6 reports Comp/Suff across $k \in \{10, \ldots, 50\}$, showing similar conclusions.

For sequential data (patient trajectories), IG (PAD-all) outperforms IG (MASK-all) on Suff in both datasets (Table 1). The IG (PAD per-trajectory) variant places the baseline point closer to cluster centers in the learned embedding space e (Fig. 1), but still lags the manifold-aware baseline on both metrics in both tasks. SIG and GIG underperformed on our MMT; For SIG, the primary factors can be aggregation and redundancy—our architecture pools token information (Transformer into GRU), so single-token perturbations (as used by SIG) often induce small output changes despite jointly important codes. For GIG, the lower scores likely reflect metric bias, Comp/Suff (insertion–deletion) favor peaky, highly concentrated saliency maps, whereas GIG's default low-gradient, noise-avoiding path yields more distributed attributions that these metrics tend to undervalue.EG, despite reduced baseline sensitivity in principle, does not surpass the IG variants here—possibly because sampling from a prior alone is insufficient for highly structured healthcare trajectories, and larger sample budgets are required to cover the overall complex structure of the patients' trajectories (see Appendix Tables 5–8 for Comp and Suff evaluated at varying k%). Quantitative results on the impact of the manifold-aware baseline for tabular (input-space) features are provided in Appendix A.8.

4.2. Impact of Path Optimization in IG

Dense, dimension-level attributions can hinder interpretability. GS-IG adds a lightweight, per-sample schedule selection to promote *token-level* sparsity without degrading faithfulness. As shown in Table 2, optimizing the schedule parameter θ via a small Bayesian search (T=10 trials) reduces the mixed $\ell_{2,1}$ group norm (sum of token ℓ_2 row norms) by roughly 5–18% across baselines, while keeping Comprehensiveness and Sufficiency essentially unchanged. This indicates that modest path adaptations can tailor explanations for improved readability at negligible cost to faithfulness.

Table 1: Faithfulness on sequential data. All IG-family methods use K=50 steps;

Method (IG variants)	MIMIC-IV		MDC	
	Comp ↑	Suff ↓	Comp ↑	Suff ↓
Gradient×Input	0.147	0.128	0.144	0.096
IG (zero baseline)	0.158	0.113	0.199	0.054
IG (pad-all baseline)	0.156	0.108	0.206	0.026
IG (mask-all baseline)	0.156	0.120	0.205	0.030
IG (pad per-trajectory baseline)	0.162	0.154	0.222	0.056
SIG	0.101	0.142	0.018	0.138
EG (10 steps)	0.142	0.149	0.166	0.057
EG (50 steps)	0.132	0.137	0.173	0.058
GIG	0.111	0.134	0.159	0.031
IG (manifold-aware; embedding space)	**0.164**	**0.078**	**0.244**	**0.022**

4.3. Qualitative Evaluation

In contrast to text and images, there is no definitive or trivial ground truth for the importance of individual clinical events in our setting, which makes it difficult to assess the effect of different baselines solely by inspection. Following guidance on falsifiable interpretability (Leavitt and Morcos, 2020), we use qualitative figures for context and pair them with *falsifiable* tests: Comprehensiveness (↑) and Sufficiency (↓)

Figure 3 contrasts three configurations on a held-out patient: (i) IG (mask-all baseline), a common practitioner choice; (ii) IG (manifold-aware baseline) (Sec. 2.2); and (iii) GS-IG (manifold-aware baseline with group sparsity; Sec. 2.3). Tokens are shown over time with signed contributions toward early death (red) or long life (blue). To preserve privacy, medical codes are anonymized as $C1$, $C2$, ...; the ICD/ATC mapping is withheld. The special token `[SEP]` marks visit boundaries (transition to the next encounter).

Across methods, the sign of token contributions is largely consistent (Fig. 3); discrepancies typically occur only when attributions are near zero rather than as genuine sign flips. This stability is expected when the path remains in a locally linear regime of a piecewise-linear network (e.g., ReLU blocks): the directional derivative of the class logit along a token-embedding axis is then approximately constant, so changing the baseline mainly rescales magnitudes without reversing directions. When sign flips do occur across baselines, they reflect IG's baseline dependence—i.e., different counterfactuals $F(x)-F(x')$ traced along different paths in embedding space (Sec. 2.2).

Furthermore, adding group sparsity (GS-IG) improves readability by pruning low-magnitude tokens while preserving faithfulness (Table 2). Additional qualitative results are provided in Appendix A.7 and A.9.

5. Related Works

Attribution methods assign importance to input components and can be categorized into three main categories: gradient-based, attention-based, and perturbation-based. Gradient methods are generally more faithful to the model (Adebayo et al., 2018; Kin-

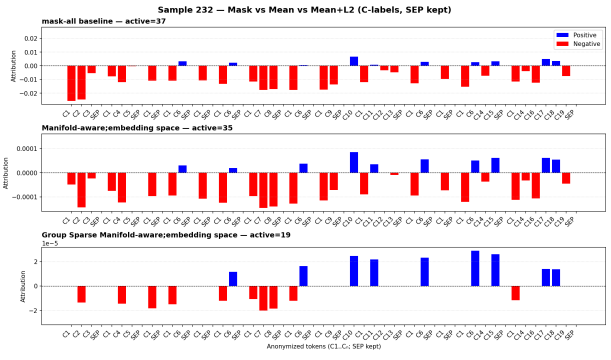


Figure 3: Qualitative comparison for a held-out patient from the MDC cohort predicted as early death. **Top:** IG with [MASK]-all baseline; **Middle:** IG with (manifold-aware) baseline; **Bottom:** GS-IG (manifold-aware + group sparsity). Each panel shows medical-code attributions over time and reports the number of *active* codes (non-zero attribution). Here, GS-IG reduces active codes from 35 to 19 (vs. the middle panel), yielding a sparser, more readable list of decisive factors. Red indicates contributions toward early death; blue indicates contributions toward long life.

Table 2: Effect of GS-IG on sequential attributions. Δ group sparsity is the relative reduction in the token-level group ℓ_2 measure. Reported Comp and Suff correspond to the optimized schedule $\theta^\star$;

.

Method (IG variants)	MIMIC-IV			MDC		
	Comp ↑	Suff ↓	$(S(\theta{=}1), S(\theta^\star))$ Δ Group sparsity (%)	Comp ↑	Suff ↓	$(S(\theta{=}1), S(\theta^\star))$ Δ Group sparsity (%)
IG (zero baseline)	0.158	0.112	(4.30,3.89) 9.5	0.185	0.051	(2.29,2.18) 5.0
IG (pad-all baseline)	0.155	0.117	(7.32,6.58) 10.2	0.206	0.025	(4.46,4.21) 5.6
IG (mask-all baseline)	0.156	0.119	(12.30,10.86) 11.6	0.204	0.029	(4.26,4.01) 6.0
IG (pad per-trajectory)	0.161	0.156	(0.60,0.55) 8.9	0.223	0.058	(0.54,0.46) 13.9
IG (manifold-aware; embedding space)	0.164	0.079	(3.01,2.46) 18.2	0.243	0.023	(3.10,2.81) 9.2

dermans et al., 2019), while attention weights may not reflect true model reliance (Jain and Wallace, 2019; Serrano and Smith, 2019; Hao et al., 2021), and perturbation approaches are often costly and sensitive to distribution shifts (Ribeiro et al., 2016b). For transformers on structured EHR, explainability has been previously explored via perturbation analyses for risk factors (Rao et al., 2022; Li et al., 2020), attention-based token/variable attribution (Lahlou et al., 2021; Rasmy et al., 2021; Avsec et al., 2021; Madan et al., 2024), SHAP adaptations (Kokalj et al., 2021), and IG for biological sequences (Madan et al., 2022).

Despite its popularity, IG is sensitive to both the baseline and the integration path. Early NLP work noted that input gradients and LIME can be effective in BERT but suffer from saturation/nonlinearity (DeYoung et al., 2019). To better respect the data manifold in embedding spaces, Sanyal and Ren (2021) proposed Discretized IG (DIG), which modifies the interpolation path so intermediate points remain meaningful in the embedding manifold. Enguehard (2023) introduced Sequential IG (SIG), computing token-level IG sequentially to reduce computational overhead for large transformers and showed it outperforms DIG. Expected Gradients (EG) averages IG over a distribution of baselines to reduce baseline sensitivity (Erion et al., 2021), and has been applied to pretrained clinical encoders (Rupp et al., 2023). Using a mean embedding as an IG baseline has been suggested conceptually in prior work, but evaluations use zero or special-token baselines instead Bastings et al., 2021.

Prior baselines for sequences often rely on special tokens (e.g., `<PAD>`/`<MASK>`), which may sit off-manifold and produce noisy interpolations. In contrast, we construct a manifold-aware baseline from empirical averages of embedding-layer outputs (and input-space means for tabular features), yielding a single, position-agnostic "null" that stays near the support of the learned embedding distribution (Sec. 2.2). This differs from DIG/SIG/EG: DIG enforces manifold-respecting *paths*, SIG restructures computation at the token level, and EG marginalizes over many baselines; our approach instead fixes a single, data-aligned baseline that removes special-token heuristics.

Several IG variants change the path between baseline and input to reduce noise or improve robustness. BlurIG (Xu et al., 2020) integrates along a blur continuum in pixel space. IG (Kapishnikov et al., 2021) makes the path adaptive by progressively advancing only low-sensitivity features. Manifold/Geodesic IG (Zaher et al., 2024) IG aligns the path to a learned data manifold, computing attributions along latent geodesics (using a VAE-induced Riemannian metric). While effective on images, these approaches introduce nontrivial engineering cost—blur pyramids and multiple forwards (BlurIG), per-step gradient ranking and path updates (GIG), and training an effective generative model for EHR trajectories plus solving geodesics (Manifold IG)—and they rely on continuous image operations that do not transfer cleanly to discrete token sequences like patient code trajectories.

Sparse explanations are easier to audit and act upon. MoreauGrad (Zhang and Farnia, 2023) smooths the classifier via a Moreau envelope and can impose ℓ_1 or group-$\ell_{2,1}$ penalties to produce (group-)sparse, robust saliency maps. Our *GS-IG* remains within the IG framework and keeps the straight geometry in embedding space and re-parameterizes only the schedule; we select θ per sample by minimizing a group-sparse $\ell_{2,1}$ objective over token embeddings, anchored by a manifold-aware baseline.

6. Conclusion

We revisited Integrated Gradients for models operating in embedding spaces and introduced two plug-and-play improvements. First, a *manifold-aware baseline*—the empirical mean of embedding-layer outputs (and input-space means for tabular features)—replaces synthetic tokens and keeps interpolates near the data manifold. Second, *GS-IG* selects a one-parameter integration schedule that minimizes a group-lasso measure of token attributions, producing concise, token-level explanations without modifying the model. Across two clinical cohorts (MIMIC-IV, MDC) and multimodal transformers, the manifold-aware baseline improved faithfulness (higher Comprehensiveness, lower Sufficiency) over common heuristics, while GS-IG reduced the mixed $\ell_{2,1}$ group norm of attributions by ≈9–18% with negligible impact on faithfulness on the manifold-aware baseline.

Efficiency and scalability. Unlike SIG (Enguehard, 2023) and DIG (Sanyal and Ren, 2021)—which incur substantial overhead from tokenwise or constrained, path-dependent computations—our manifold-aware baseline is computed once, cost essentially unchanged. GS-IG adds only a lightweight, scalar schedule search (small T) on top of standard IG. In practice, this makes our approach a scalable, efficient alternative that is particularly well-suited for large MMT architectures.

Limitations The baseline depends on the validation distribution and may require recalibration under data shift. For inputs near the mean, completeness yields small $|F(x) - F(x')|$ (low-contrast); we accept this trade-off to avoid Out-of-distribution (OOD) paths, and our faithfulness metrics show the net benefit. The schedule search optimizes a single scalar; richer schedules or joint selection with sparsity thresholds could further improve compactness.

Future works Considering the learned data manifold in embedding space, by methods like data density in those regions, when optimizing the sparsity objective, can be important. Beyond intrinsic metrics, human-in-the-loop studies (e.g., clinician audits), additional data modalities, and causal faithfulness evaluations are important next steps.

Acknowledgments

We thank Jonas Björk and Olle Melander for facilitating access to the data and for their valuable guidance in understanding and interpreting the dataset and the tasks. We also thank Jens Lundstrom for valuable feedback.

This study was conducted as part of the AIR Lund (Artificially Intelligent use of Registers at Lund University) research environment and was funded by the Swedish Research Council (VR, grant 2019-00198). Additional support was provided by CAISR Health, funded by the Knowledge Foundation (KK-stiftelsen) in Sweden (grant 20200208 01 H).

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

Ali Amirahmadi, Farzaneh Etminani, Jonas Björk, Olle Melander, and Mattias Ohlsson. Trajectory-ordered objectives for self-supervised representation learning of temporal healthcare data using transformers: Model development and evaluation study. *JMIR Medical Informatics*, 13(1):e68138, 2025a.

Ali Amirahmadi, Farzaneh Etminani, and Mattias Ohlsson. Adaptive noise-augmented attention for enhancing transformer fine-tuning on longitudinal medical data. *Frontiers in Artificial Intelligence*, 8:1663484, 2025b.

Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10): 1196–1203, 2021.

Jasmijn Bastings, Sebastian Ebert, Polina Zablotskaia, Anders Sandholm, and Katja Filippova. " will you find these shortcuts?" a protocol for evaluating the faithfulness of input salience methods for text classification. *arXiv preprint arXiv:2111.07367*, 2021.

G Berglund, S Elmståhl, L Janzon, and SA Larsson. Design and feasibility. *Journal of internal medicine*, 233(1):45–51, 1993.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx,

Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. Eraser: A benchmark to evaluate rationalized nlp models. *arXiv preprint arXiv:1911.03429*, 2019.

Joseph Enguehard. Sequential integrated gradients: a simple but effective method for explaining language models. *arXiv preprint arXiv:2305.15853*, 2023.

Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott M Lundberg, and Su-In Lee. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature machine intelligence*, 3(7):620–631, 2021.

Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3681–3688, 2019.

Lin Lawrence Guo, Ethan Steinberg, Scott Lanyon Fleming, Jose Posada, Joshua Lemmon, Stephen R Pfohl, Nigam Shah, Jason Fries, and Lillian Sung. Ehr foundation models improve robustness in the presence of temporal distribution shift. *Scientific Reports*, 13(1):3767, 2023.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. Self-attention attribution: Interpreting information interactions inside transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12963–12971, 2021.

Juyeon Heo, Sunghwan Joo, and Taesup Moon. Fooling neural network interpretations via adversarial model manipulation. *Advances in neural information processing systems*, 32, 2019.

Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv. *PhysioNet. Available online at: https://physionet.org/content/mimiciv/1.0/(accessed August 23, 2021)*, 2020.

Andrei Kapishnikov, Subhashini Venugopalan, Besim Avci, Ben Wedin, Michael Terry, and Tolga Bolukbasi. Guided integrated gradients: An adaptive path method for removing noise. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5050–5058, 2021.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un) reliability of saliency methods. *Explainable AI: Interpreting, explaining and visualizing deep learning*, pages 267–280, 2019.

Enja Kokalj, Blaž Škrlj, Nada Lavrač, Senja Pollak, and Marko Robnik-Šikonja. Bert meets shapley: Extending shap explanations to transformer-based classifiers. In *Proceedings of the EACL hackashop on news media content analysis and automated report generation*, pages 16–21, 2021.

Chuhong Lahlou, Ancil Crayton, Caroline Trier, and Evan Willett. Explainable health risk predictor with transformer-based medicare claim encoder. *arXiv preprint arXiv:2105.09428*, 2021.

Matthew L Leavitt and Ari Morcos. Towards falsifiable interpretability research. *arXiv preprint arXiv:2010.12016*, 2020.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):7155, 2020.

Scott Lundberg. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

Sumit Madan, Victoria Demina, Marcus Stapf, Oliver Ernst, and Holger Fröhlich. Accurate prediction of virus-host protein-protein interactions via a siamese neural network using deep protein sequence embeddings. *Patterns*, 3(9), 2022.

Sumit Madan, Manuel Lentzen, Johannes Brandt, Daniel Rueckert, Martin Hofmann-Apitius, and Holger Fröhlich. Transformer models in biomedicine. *BMC Medical Informatics and Decision Making*, 24(1):214, 2024.

Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and trends® in Optimization*, 1(3):127–239, 2014.

Shishir Rao, Yikuan Li, Rema Ramakrishnan, Abdelaali Hassaine, Dexter Canoy, John Cleland, Thomas Lukasiewicz, Gholamreza Salimi-Khorshidi, and Kazem Rahimi. An explainable transformer-based deep learning model for the prediction of incident heart failure. *ieee journal of biomedical and health informatics*, 26(7):3362–3372, 2022.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86, 2021.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016a.

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*, pages 1135–1144, 2016b.

Maurice Rupp, Oriane Peter, and Thirupathi Pattipaka. Exbehrt: Extended transformer for electronic health records. In *International Workshop on Trustworthy Machine Learning for Healthcare*, pages 73–84. Springer, 2023.

Soumya Sanyal and Xiang Ren. Discretized integrated gradients for explaining language models. *arXiv preprint arXiv:2108.13654*, 2021.

Sofia Serrano and Noah A Smith. Is attention interpretable? *arXiv preprint arXiv:1906.03731*, 2019.

Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*, 2016.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMlR, 2017.

Noah Simon and Robert Tibshirani. Standardization and the group lasso penalty. *Statistica Sinica*, 22 (3):983, 2012.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A Pfeffer, Jason Fries, and Nigam H Shah. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6(1):135, 2023.

Shawn Xu, Subhashini Venugopalan, and Mukund Sundararajan. Attribution in scale and space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9680–9689, 2020.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67, 2006.

Eslam Zaher, Maciej Trzaskowski, Quan Nguyen, and Fred Roosta. Manifold integrated gradients: Riemannian geometry for feature attribution. *arXiv preprint arXiv:2405.09800*, 2024.

Jingwei Zhang and Farzan Farnia. Moreaugrad: Sparse and robust interpretation of neural networks via moreau envelope. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2021–2030, 2023.

Appendix A. Technical Appendices and Supplementary Material

A.1. Details on Group Lasso for token sparsity

Why this is Group Lasso. In Eq. (5) we penalize the rowwise ℓ_2 norms of the embedding-space IG

matrix:

$$\mathcal{L}(\theta) = \lambda_{\text{grp}} \sum_{j=1}^{L} m_j \left\| \widehat{\text{IG}}_{\text{emb}}[j,:] \right\|_2 = \lambda_{\text{grp}} \left\| \widehat{\text{IG}}_{\text{emb}} \right\|_{2,1} \quad \text{with groups } S_j = \{(j, 1{:}d_e)\}. \tag{6}$$

which is the group lasso (a mixed $\ell_{2,1}$) penalty applied to token groups S_j. Each group is an entire token embedding, so the penalty treats all d_e coordinates of a token symmetrically and only depends on their Euclidean norm. This rotational invariance within a group ensures that, if a token is selected, GS-IG does not favor any particular embedding dimension; if it is not, the *entire* row is suppressed.

How group lasso induces token sparsity. Unlike an elementwise ℓ_1 penalty, the $\ell_{2,1}$ penalty is nondifferentiable at the origin in each group. This geometry yields an "all-in / all-out" effect: small groups are set *exactly* to zero, while large groups are uniformly shrunk. Formally, for an auxiliary variable $v \in \mathbb{R}^{d_e}$ (one token's row), the proximal operator of the group penalty,

$$\text{prox}_{\alpha\|\cdot\|_2}(v) = \arg\min_{z \in \mathbb{R}^{d_e}} \tfrac{1}{2}\|z - v\|_2^2 + \alpha\|z\|_2, \tag{7}$$

admits the closed-form *group soft-thresholding* (GST):

$$\text{GST}_\alpha(v) = \begin{cases} \mathbf{0}, & \|v\|_2 \le \alpha, \\ \left(1 - \frac{\alpha}{\|v\|_2}\right) v, & \|v\|_2 > \alpha, \end{cases} \tag{8}$$

so any group whose norm falls below the threshold collapses to the exact zero vector Yuan and Lin (2006); Parikh et al. (2014). MoreauGrad explicitly derives and uses this GST operator when enforcing group sparsity inside its optimization loop, and shows that increasing the group-sparsity coefficient removes more groups while preserving desirable smoothness/robustness properties (Zhang and Farnia, 2023, Def. 4; GST formula; robustness Thm. 2). In our case, we do not solve a proximal subproblem; instead, we *select* the path schedule parameter θ that *minimizes* the same $\ell_{2,1}$ measure of the IG output. Because each active token incurs an additive cost proportional to its row-norm, the minimizer $\theta^\star$ naturally concentrates attribution on *fewer* token rows (those with the largest aggregate effect) and suppresses the rest.

Intuition in our setting. Completeness fixes the *total signed attribution* $\sum_i \text{IG}_i = F(x) - F(x')$ for any smooth path between x' and x. The freedom in θ redistributes that fixed "budget" across tokens. The $\ell_{2,1}$ objective makes dispersion expensive: spreading small amounts of attribution over many tokens increases $\sum_j \|\cdot\|_2$, whereas concentrating it on a handful of tokens reduces the sum. Thus, selecting $\theta^\star = \arg\min_\theta \|\widehat{\text{IG}}_{\text{emb}}(x, x'; \theta)\|_{2,1}$ yields *token-sparse* explanations without modifying the model or its loss.

Practical notes. All token groups share size d_e, so no group-size reweighting is needed. If group sizes differed (e.g., heterogeneous modalities), one can scale each group's term by a weight w_j (commonly $w_j = \sqrt{|S_j|}$) to avoid bias toward smaller groups Simon and Tibshirani (2012).

A.2. Models Architecture

Our model is designed to integrate longitudinal medical histories and non-temporal structured data to predict early mortality in patients, see Figure 4. The architecture consists of three main components: a transformer-based encoder for sequential medical encounters, a feedforward network for structured tabular data, and a cross-attention mechanism for modality fusion. This design enables the model to effectively capture both temporal dependencies and non-temporal structured patient characteristics.

The first component processes patient medical histories, which consist of diagnoses and prescribed medications recorded across multiple hospital visits. Each encounter is represented as a set of medical codes, and after preprocessing, medical tokens are organized into a sequence where special separator tokens delineate individual visits (medical codes within a visit have the positions) following Li et al. (2020); Rasmy et al. (2021). To encode meaningful representations of these sequences, we utilize TOO-BERT Amirahmadi et al. (2025a), a pretrained transformer model specialized for patient trajectories, which is trained with masked language modeling and trajectory order prediction objectives. TOO-BERT captures contextual dependencies within medical sequences, leveraging self-attention mechanisms to model interactions across different visits and different diagnoses and medications. To capture the higher-level trajectory representation, we introduce a GRU layer on top of the TOO-BERT encoder, which aggregates medical codes across a trajectory and outputs the final hid-

den state as the patient's overall trajectory representation.

The second component of the model processes non-temporal structured data. Categorical variables are first projected into continuous space using the $M(z_i)$ mapping function, then concatenated with numerical features before being passed through two fully connected layers with ReLU activation. This component learns a dense representation of non-temporal structured patient attributes, providing complementary information to the temporal medical history.

The third component integrates the two modalities using a cross-attention mechanism. The representation of the medical trajectory (TOO-BERT output) serves as the query, while the structured tabular representation is used as the key and value. This mechanism allows the model to dynamically attend to the most relevant non-temporal structured features in the context of a patient's medical history, learning meaningful interactions between medical trajectories and structured data. The cross-attended representation is then passed through a final feedforward network, followed by a sigmoid activation function to compute the probability of early death.

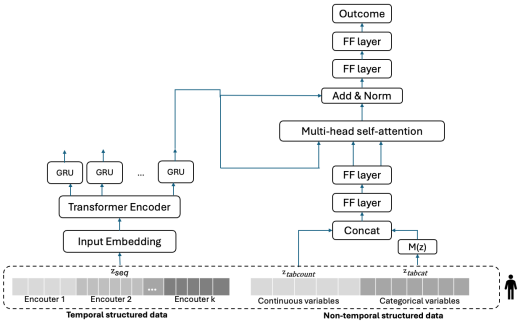


Figure 4: Multi-modal Transformer (MMT) architecture. The model integrates sequential medical records and phenotypic features, processing temporal data with a transformer encoder and tabular data with fully connected layers. A cross-attention mechanism fuses both modalities.

A.2.1. Multi-modal Transformer performance

We first trained single-modality models, then fused modalities. Using only tabular data collected at enrollment, an MLP achieved AUC 0.788 for early death and 0.661 for HF prediction. Using only EHR trajectories, the TOO-BERT–based model reached AUC 0.932 (early death) and 0.877 (HF). Combining structured tabular data with longitudinal EHR data in the MMT further improved performance to AUC 0.953 and 0.902 for early death and HF, respectively, demonstrating the benefit of using both modalities.

A.3. Dataset Specifications

In our study, we utilized two cohorts: the Malmö Diet and Cancer (MDC) study (Berglund et al., 1993), and the MIMIC-IV hospital (hosp) module (Johnson et al., 2020). Both datasets are split into 75% train, 10% validation, and 15% test.

MDC. The MDC dataset contains electronic health record (EHR) trajectories for a general-population cohort, including histories of diagnoses and medications recorded using ICD and ATC codes, together with rich phenotypic and lifestyle data (e.g., diet, heredity, socio-economic factors, lifestyle, occupation). The tabular component comprises 912 features collected at enrollment. The cohort includes 30,000 individuals with 531,000 recorded visits spanning 1992–2020. After preprocessing, 16% of participants are labeled as early-death cases.

MIMIC-IV (hosp). The MIMIC-IV hosp module contains inpatient EHR trajectories for approximately 173,000 patients across 407,000 hospital visits from 2008–2019, with about 10.6 million medical codes (diagnoses and medications). We also use available demographics (gender, race, marital status, insurance, language, and age at last encounter). After preprocessing, 36% of patients are labeled as first-time *heart failure* diagnoses.

Tables 3 and 4 provide detailed statistics on the two datasets used in this study before preprocessing.

Table 3: Characteristics of longitudinal EHR trajectory data

Dataset	MDC	MIMIC-IV
#Patients	30 K	173 K
#Visits	531 K	408 K
#All Medical Codes	7.6 M	10.6 M

Table 4: Characteristics of the MDC recorded phenotype data

#Patients	30 K
#Nominal features	606
#Ordinal features	186
#Numerical features	120
#All features	912

A.4. Problem Formulation

Each dataset D consists of a set of patients P, formally defined as:

$$D = \{P^1, P^2, \ldots, P^{|D|}\}. \tag{9}$$

For each patient P^i, we have a combination of longitudinal EHR data—including diagnoses and prescribed medications—and non-temporal structured phenotype features. Each patient is represented by a sequence of medical encounters and a set of structured phenotype features as:

$$P^i = \left[\{V_1^i, V_2^i, \ldots, V_O^i\}, \{\text{phen}_1, ..., \text{phen}_{912}\}\right], \tag{10}$$

where O is the total number of recorded visits for patient i, and phen_j represents the structured phenotype information. Each visit V_j^i contains the set of diagnosis codes and prescribed medications, defined as:

$$V_j^i = I_j \cup M_j, \tag{11}$$

where $I_j \subset I$ represents the diagnosis codes (ICD) and $M_j \subset M$ represents the prescribed medications (ATC) recorded at visit V_j^i. To reduce sparsity, we exclude infrequently occurring medical codes and retain only the first four digits of ICD and ATC codes to ensure meaningful aggregation.

To focus on relevant predictive patterns and avoid label leakage, we censor recent records near the outcome. In MDC, we exclude all medical codes recorded within the four years preceding the index event. For early-death cases, this removes codes from the four years prior to death; for long-lived controls, we remove codes from the four years before the last recorded visit (or date of death) to ensure temporal alignment. Similarly, for heart-failure prediction in MIMIC-IV, we exclude heart-failure–specific medication codes and omit information from the final encounter (last visit) so that peri-diagnostic signals do not leak into training or evaluation.

To help the model capture the temporal structure of patient trajectories, we introduce special tokens, inspired by BERT-like architectures. A `[CLS]` token is inserted at the beginning of each patient's trajectory to represent global sequence-level embeddings, while a `[SEP]` token is placed between visits to distinguish encounters. For all experiments, we consider at most the last 200 medical codes per patient; shorter sequences are right-padded with `[PAD]` tokens.Consequently, the structured representation of a patient trajectory is given as:

$$P^i = \big[\,\{[\texttt{CLS}],\, V_1^i,\, [\texttt{SEP}],\, V_2^i,\, [\texttt{SEP}],\, \ldots,\, V_O^i,\, [\texttt{SEP}]\},\; \{\text{phen}_1,\, \ldots,\, \text{phen}_{912}\}\,\big]. \tag{12}$$

which provides valuable contextual cues to the model for learning meaningful temporal relationships.

The goal is to predict the probability of an adverse event—premature death in MDC and first-time heart failure in MIMIC-IV—given a patient's longitudinal history and structured phenotype data. Formally, for patient i we model

$$p_\theta(e \mid P^i), \qquad e \in \{\text{early-death},\ \text{heart-failure}\},$$

where P^i denotes the combined trajectory and phenotypic features, and p_θ is the output of the trained model.

A.5. Evaluation Metrics

We evaluate the faithfulness of model explanations with Comprehensiveness (Comp) and Sufficiency (Suff), two widely used metrics in explainability research (DeYoung et al., 2019; Sanyal and Ren, 2021; Enguehard, 2023).

Comp measures the drop in model performance when the most $k\%$ important features (as determined by an attribution method) are removed, where a greater drop indicates a more faithful explanation. For sequential inputs (temporal structured data), we excluded special tokens and removed the top $k\%$ most attributed tokens from the sequence while preserving the structure of patient histories, including the number of visits. For tabular features (both numerical and categorical data), direct removal is not feasible. Instead, we replaced the top $k\%$ most attributed feature values with values randomly sampled from the other samples. To account for variability, this process was repeated 20 times with different random samples,

and the mean Comp score was reported. Formally, Comp can be presented as :

$$\text{Comprehensiveness} = F(x_p)_j - F(x_p \backslash r_p)_j \qquad (13)$$

where $F(x_p)_j$ represents the model's original prediction score for class j, patient p, and r_p denotes the $k\%$ most important features (or tokens) identified by the attribution method. We set $k = 20\%$ for Comprehensiveness in all experiments.

Conversely, Suff evaluates how well the model retains its confidence when only the most important features are retained, with a smaller drop suggesting that the model relies primarily on these features for decision-making. For tabular data, similar to Comp, we replaced the values of the lowest attributed $(1 - k\%)$ features 20 times and computed the mean Suff score. The metric is formally defined as:

$$\text{Sufficiency} = F(x_p)_j - F(r_p)_j \qquad (14)$$

where r_p represents the retained top $k\%$ most important features. We used $k = 10\%$ for Suff to better capture the model's reliance on only highly attributed features.

A.6. Comprehensiveness and Sufficiency Across $k\%$

Setup. In the main paper we report Comp and Suff at a fixed top-$k\%$ of tokens/features. Here we provide the same metrics across multiple k values (10–50%) for two datasets (MIMIC-IV, MDC) and three IG variants: IG with the mean-embedding ("manifold-aware") baseline, IG with `[MASK]`-all, and IG with a zero baseline. Across both datasets, the mean-embedding baseline outperforms the alternatives for all k values, with the clearest gains in the low-k regime that is most relevant for concise review.

A.7. Qualitative Evaluation on Patient Trajectory (Sequential) Data

Figures 5–8 present additional held-out patient trajectories. Each figure is organized vertically: **Top** row: IG with [MASK]-all baseline; **Middle** row: IG with the manifold-aware baseline; **Bottom** row: GS-IG (same manifold-aware baseline + sparsity schedule). The middle and bottom rows, therefore, differ only by the group-sparsity schedule.

Each panel displays token-level attributions over time and reports the number of *active medical codes* (non-zero attribution) for that patient. Across cases, GS-IG typically produces *fewer active codes* and clearer summaries, yielding sparser and more readable results for domain users.

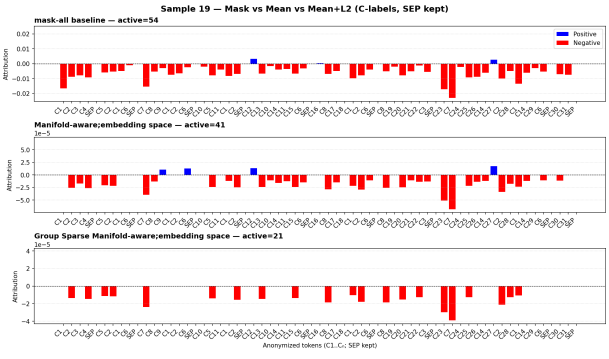


Figure 5: Held-out patient, vertical layout: **Top** IG ([MASK]-all), **Middle** IG (manifold-aware), **Bottom** GS-IG (manifold-aware + group sparsity). Each panel shows token-level attributions over time and reports the number of active codes (non-zero attributions). Here, GS-IG reduces active codes from 41 to 21 (vs. the middle panel), yielding a sparser, more readable list of decisive factors. Red = contribution toward early-death, blue = contribution toward long-life. Codes anonymized as C_n; `[SEP]` marks visits.

A.8. Quantitative Evaluation for the Tabular Data

For phenotypic tabular data, the empirical-mean (input-space) baseline also provides the most faithful attributions, substantially outperforming Gradient×Input and IG with a zero baseline (Table 9). This suggests that a simple empirical-mean baseline is a reasonable null for non-embedded features as well. The difference in the scale of Comp/Suff between MIMIC-IV and MDC largely reflects feature richness: in MIMIC-IV we have only six demographic variables, which provide limited signal for HF prediction (MLP AUC $\approx$ 0.66 when used alone), whereas MDC includes 912 phenotypic variables that are considerably more informative (MLP AUC $\approx$ 0.79) (A.2.1).

Table 5: MIMIC-IV — Comprehensiveness (higher is better) across $k\%$ of tokens retained.

$k\%$	10	20	30	40	50
IG (mean-embedding)	0.141	0.164	0.182	0.203	0.228
IG (`[MASK]`-all)	0.125	0.156	0.180	0.202	0.228
IG (zero baseline)	0.131	0.158	0.178	0.203	0.228

Table 6: MIMIC-IV — Sufficiency (lower is better) across $k\%$ of tokens retained.

$k\%$	10	20	30	40	50
IG (mean-embedding)	0.078	0.007	-0.012	-0.019	-0.023
IG (`[MASK]`-all)	0.120	0.066	0.030	0.007	0.000
IG (zero baseline)	0.113	0.053	0.036	0.029	0.026

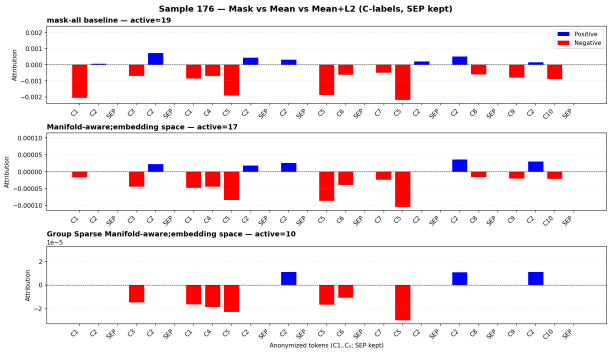


Figure 6: Held-out patient, vertical layout: **Top** IG ([MASK]-all), **Middle** IG (manifold-aware), **Bottom** GS-IG (manifold-aware + group sparsity). Each panel shows token-level attributions over time and reports the number of active codes (non-zero attributions). Here, GS-IG reduces active codes from 17 to 10 (vs. the middle panel), yielding a sparser, more readable list of decisive factors. Red = contribution toward early-death, blue = contribution toward long-life. Codes anonymized as C_n; `[SEP]` marks visits.

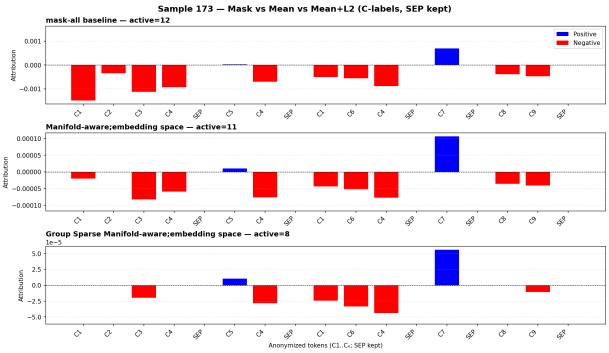


Figure 7: Held-out patient, vertical layout: **Top** IG ([MASK]-all), **Middle** IG (manifold-aware), **Bottom** GS-IG (manifold-aware + group sparsity). Each panel shows token-level attributions over time and reports the number of active codes (non-zero attributions). Here, GS-IG reduces active codes from 11 to 8 (vs. the middle panel), yielding a sparser, more readable list of decisive factors. Red = contribution toward early-death, blue = contribution toward long-life. Codes anonymized as C_n; `[SEP]` marks visits.

Table 7: MDC — Comprehensiveness (higher is better) across $k\%$ of tokens retained.

$k\%$	10	20	30	40	50
IG (mean-embedding)	0.132	0.244	0.330	0.367	0.332
IG (`[MASK]`-all)	0.107	0.205	0.277	0.274	0.249
IG (zero baseline)	0.111	0.199	0.268	0.261	0.238

Table 8: MDC — Sufficiency (lower is better) across $k\%$ of tokens retained.

$k\%$	10	20	30	40	50
IG (mean-embedding)	0.022	-0.010	-0.017	-0.021	-0.023
IG (`[MASK]`-all)	0.030	-0.009	-0.017	-0.020	-0.023
IG (zero baseline)	0.054	0.020	0.009	-0.002	-0.011

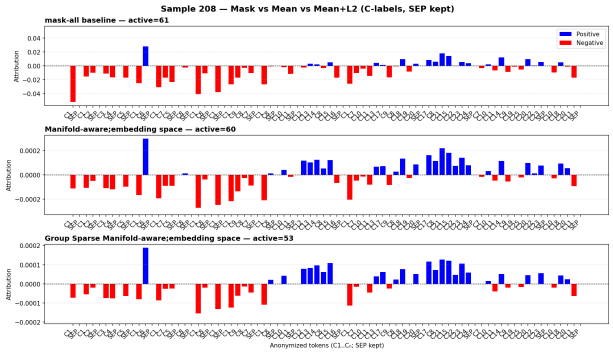


Figure 8: Held-out patient, vertical layout: **Top** IG ([MASK]-all), **Middle** IG (manifold-aware), **Bottom** GS-IG (manifold-aware + group sparsity). Each panel shows token-level attributions over time and reports the number of active codes (non-zero attributions). Here, GS-IG reduces active codes from 60 to 53 (vs. the middle panel), yielding a sparser, more readable list of decisive factors. Red = contribution toward early-death, blue = contribution toward long-life. Codes anonymized as C_n; `[SEP]` marks visits.

Table 9: Faithfulness on phenotypic (tabular) data: comprehensiveness (Comp; ↑) and sufficiency (Suff; ↓). The manifold-aware baseline here is the empirical mean in input space (Sec. 2.2).

Method (IG variants)	**MIMIC-IV**		**MDC**	
	Comp ↑	**Suff ↓**	**Comp ↑**	**Suff ↓**
Gradient×Input	0.005	0.006	0.326	0.122
IG (zero baseline)	0.005	0.007	0.327	0.123
IG (manifold-aware; input space)	**0.011**	**0.001**	**0.843**	**0.027**

A.9. Qualitative Evaluation for the Tabular Data

Figure 9 compares tabular features with the largest absolute attributions under IG with a manifold-aware baseline versus a zero baseline for a representative patient. Entries marked [MA] denote features selected by the manifold-aware baseline; [Z] denotes the zero-baseline selection. Here, the zero baseline concentrates on sex and sex-related variables (prefix `kv`), suggesting a baseline-driven artifact. In contrast, the manifold-aware baseline highlights a more diverse and clinically plausible set of contributors, indicating that a realistic reference improves attribution specificity.

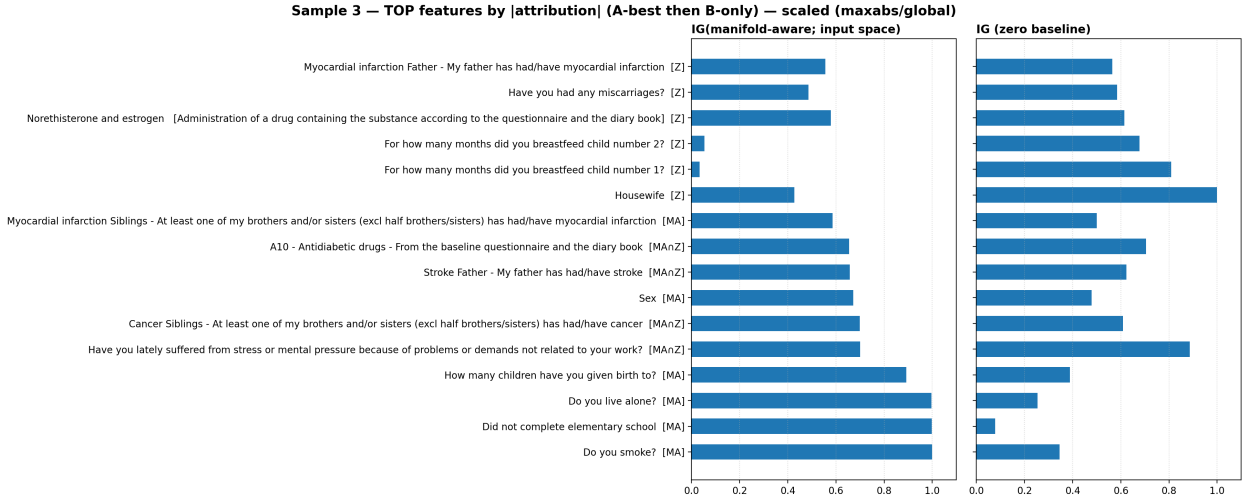


Figure 9: Tabular features with the highest absolute attribution under IG using a manifold-aware baseline (left) versus a zero baseline (right). The bracket tag indicates the originating baseline: `[MA]` for manifold-aware, `[Z]` for zero.

A.10. Empirical Evidence That the Mean-Embedding Baseline Resides in Higher-Support Regions and Improves IG Path Quality

Setup. Let $\mathbf{x}$ denote a sequence and let $\boldsymbol{\phi}(\mathbf{x}) \in \mathbb{R}^d$ be the embedding-layer output flattened to a sequence-embedding vector. For each baseline $b \in \{\texttt{zero}, \texttt{pad-all}, \texttt{mask-all}, \texttt{mean-emb}\}$, we form a single baseline vector $\mathbf{z}_b \in \mathbb{R}^d$ in the same space. Let $\mathcal{Z}_{\text{val}} = \{\boldsymbol{\phi}(\mathbf{x}_i)\}_{i=1}^N$ be the validation *sequence* cloud in the representation space where integrated gradients (IG) operates.

Goal. Quantify whether a baseline lies in a high-support neighborhood of the *sequence-embedding* distribution, and thereby whether IG paths from that baseline are less exposed to off-distribution (OOD) regions.

A.10.1. Metrics in Sequence-Embedding Space

We report three complementary measures, all computed on sequence-embedding vectors:

1. **Average Euclidean distance to validation sequences**
2. **Average Mahalanobis distance to validation sequences**
3. **KDE log-density under a Gaussian KDE fitted on a whitened cloud**: We whiten $\mathcal{Z}_{\text{val}}$ to zero mean and identity covariance via $\tilde{\mathbf{z}} = \mathbf{W}(\mathbf{z} - \boldsymbol{\mu})$ with $\mathbf{W}\boldsymbol{\Sigma}\mathbf{W}^\top = \mathbf{I}$, fit an isotropic Gaussian KDE p_{KDE} on $\{\tilde{\mathbf{z}}_i\}_{i=1}^N$, and report

$$\text{KDELog}(\mathbf{z}_b) \;=\; \log p_{\text{KDE}}(\mathbf{W}(\mathbf{z}_b - \boldsymbol{\mu}))\,.$$

Smaller is better for AvgEuc and Mah; larger (less negative) is better for KDELog.

A.10.2. Emprical Results

Table 10: MDC (early death): baseline support in sequence-embedding space

Baseline	AvgEuc ↓	Mah ↓	KDELog ↑
`zero`	62.83	11135.45	-6.20×10^{8}
`pad-all`	69.94	17.62	-5.92×10^{3}
`mask-all`	69.96	109.15	-1.21×10^{4}
`mean-emb`	**45.01**	**14.83**	-5.90×10^{3}

Across both tasks, the `mean-emb` baseline is closest to the empirical sequence distribution (lowest

Table 11: MIMIC-IV (heart failure): baseline support in sequence-embedding space

Baseline	AvgEuc ↓	Mah ↓	KDELog ↑
`zero`	75.67	18957.76	-2.81×10^{9}
`pad-all`	80.03	3529.01	-1.13×10^{7}
`mask-all`	80.04	3543.21	-1.14×10^{7}
`mean-emb`	**37.60**	**10.20**	-5.93×10^{3}

AvgEuc and Mah) and attains the highest (least negative) KDELog. The `zero` baseline is catastrophically OOD by the metrics.

A.10.3. Why These Metrics Speak to IG Path Quality

Integrated gradients (IG) computes attributions by integrating gradients along a path in *embedding space*, from baseline $\mathbf{z}_b$ to input $\boldsymbol{\phi}(\mathbf{x})$.

Low-density regions correspond to parts of representation space where the model is weakly constrained by data and gradients are less stable. Placing $\mathbf{z}_b$ in a high-support neighborhood (as quantified by AvgEuc, Mah, KDELog) reduces immediate exposure to OOD areas, yielding smoother gradient accumulation—consistent with the observed faithfulness gains.

We do not claim a global guarantee across all inputs in a multi-modal cloud. Two observations mitigate this concern: (i) *Starting point matters*: off-distribution baselines (e.g., `zero`, `mask-all`) force the path to immediately traverse unsupported regions; `mean-emb` measurably avoids this initial excursion (smaller AvgEuc/Mah, larger KDELog). (ii) *Global diagnostics*: Mahalanobis distance and whitened-KDE density summarize typicality across modes; the mean-emb baseline's values indicate it sits in a high-support neighborhood.

Limitations and Scope. Our claim is empirical: starting the IG path at `mean-emb` places it in a markedly higher-support region and, in practice on two EHR tasks, correlates with more stable and faithful attributions. We do not assert a theoretical guarantee that paths avoid all low-density regions in multi-modal settings.

A.11. High-level Background: Post-hoc Explainability and Integrated Gradients

Post-hoc explainability. Post-hoc methods explain a trained model's predictions without changing the model itself. Prominent families include local surrogate methods (e.g., LIME) and additive game-theoretic approaches (e.g., SHAP), alongside gradient-based saliency methods for deep networks (Ribeiro et al., 2016a; Lundberg, 2017).

Gradient saliency and the saturation challenge. The simplest gradient explanation scores each feature by the partial derivative $\partial F(x)/\partial x_i$ (or Gradient×Input). However, deep models can exhibit *saturation* (near-flat regions) around confident predictions, yielding tiny local gradients even for decisive features and thus underestimating importance (Sundararajan et al., 2017).

Integrated Gradients (IG): intuition, axioms, and mechanics. IG combats saturation by accumulating gradients *along a path* from a *baseline* x' to the input x. For a smooth path $\gamma : [0,1] \to \mathbb{R}^d$ with $\gamma(0) = x'$, $\gamma(1) = x$, the attribution to feature i is

$$\mathrm{IG}_i(x; x') \;=\; \int_0^1 \frac{\partial F(\gamma(t))}{\partial \gamma_i(t)} \frac{\mathrm{d}\gamma_i(t)}{\mathrm{d}t}\,\mathrm{d}t, \qquad (15)$$

approximated by a Riemann sum in practice. IG satisfies key axioms that ground its use: Implementation Invariance (functionally equivalent models yield identical attributions), Sensitivity(if two inputs differ in exactly one feature and their outputs differ, that feature must receive nonzero attribution) and Completeness (attributions sum to $F(x) - F(x')$, i.e., total credit is conserved) (Sundararajan et al., 2017). These properties distinguish IG from many heuristic saliency maps.

Baselines and paths: the two design levers. IG requires (i) a baseline x' that plays the role of a counterfactual reference (e.g., a black image), and (ii) an integration path γ between x' and x. The default is the straight line $\gamma(t) = x' + t(x - x')$. In all cases, the baseline controls the counterfactual contrast $F(x) - F(x')$, whereas the path (or its parameterization) governs where and how gradients are sampled along the way.