Explaining RhythmFormer: A Systematic XAI Analysis of Periodic Sparse Attention for Remote Photoplethysmography

Louis Chen[1], Torbjörn E. M. Nordling[1,*]

[1]*Department of Mechanical Engineering, National Cheng Kung University, Tainan, Taiwan*

Abstract

Remote photoplethysmography (rPPG) transformers achieve low heart-rate error on benchmarks, yet their decisions remain opaque—a growing concern as rPPG moves toward clinical heart rate estimation. Existing rPPG XAI is dominated by qualitative heatmap inspection without quantitative faithfulness metrics or physiology-grounded validation, leaving a gap between visual plausibility and auditable evidence. We address this gap with three contributions. First, we adapt four attribution methods (raw attention, rollout, flow, Beyond Intuition) to RhythmFormer's bi-level routing attention with top-k selection. Second, we introduce a skin coverage metric quantifying how much attribution mass falls on physiologically valid skin regions. Third, we adapt the SaCo faithfulness coefficient from its original classification setting to rPPG regression by using the MAE between original and perturbed predicted rPPG waveforms as the perturbation impact. Applying these tools, we quantify a multi-hop leakage effect under sparse top-k routing: attention rollout almost completely restores the connections that individual refined-attention layers explicitly set to zero, while attention flow is exposed to the same path-based caveat. Beyond Intuition mitigates this via its value-projection-weighted rollout and gradient-supported mask, attaining the highest median refined skin coverage (0.826 vs. 0.567 for vanilla rollout) and faithfulness (F=0.917) among the evaluated methods on UBFC-rPPG with RhythmFormer. We therefore identify Beyond Intuition as the strongest candidate in this setting, while emphasizing that validation across diverse datasets and model variants remains an important direction for future work. A case study on a low-SaCo outlier further shows all four methods recovering consistently once an artefactual region is replaced, suggesting consistent SaCo behavior across attribution families in this illustrative case. Together, these two quantitative metrics move XAI for rPPG beyond visual heatmap inspection toward auditable numerical evidence about spatial alignment and perturbation faithfulness—a step toward more trustworthy rPPG XAI.

Keywords

Explainable AI, remote photoplethysmography, attention attribution, faithfulness evaluation

1. Introduction

Remote photoplethysmography (rPPG) estimates vital signs from subtle skin color changes in ordinary video, avoiding the discomfort, infection risk, and deployment limits of contact electrocardiography (ECG) and photoplethysmography (PPG) in settings such as neonatal care and remote monitoring [1]. Deep learning has moved rPPG from signal-processing pipelines [2, 3] to 2-dimensional convolutional neural networks (2D CNNs) [4, 5], 3D CNNs [6], and transformers [7, 8, 9, 10], but robustness to illumination, motion, and skin tone variation remains difficult. RhythmFormer [9] is a useful target for explainable artificial intelligence (XAI) because it reports competitive benchmark performance with a single-stream architecture, avoiding auxiliary branches or multi-task heads that complicate attribution. As rPPG transformers approach medical use, explanations should test whether model focus is physiologically plausible, for example whether it emphasizes facial regions known to carry strong plethysmographic information such as the forehead and cheeks [11]. Current rPPG XAI studies usually inspect one explanation family at a time, including raw attention [10], Grad-CAM [12], or gradient

EXPLIMED 2026 - Third Workshop on Explainable Artificial Intelligence for the medical domain - 15-17 August 2026, Bremen, Germany

*Corresponding author.

louis.chen@nordlinglab.org (L. Chen); torbjorn.nordling@nordlinglab.org (T. E. M. Nordling)

https://nordlinglab.org (L. Chen)

0009-0001-8373-2295 (L. Chen); 0000-0003-4867-6707 (T. E. M. Nordling)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-26

CEUR-WS.org/Vol-4339/paper_3.pdf

saliency [13, 14], and rarely quantify faithfulness. Raw attention is also contested as an explanation proxy [15, 16, 17]: a plausible heatmap may not identify features that causally affect the prediction.

In this work, we present a comprehensive XAI framework for analyzing RhythmFormer's periodic sparse attention. Our contributions are:

- We adapt raw attention, attention rollout [18], attention flow [18], and Beyond Intuition [19] to RhythmFormer's sparse top-k bi-level routing attention.
- We define skin coverage, measuring attribution mass on facial skin, both eyebrows, the nose, and the neck across Hierarchical Temporal Periodic Transformer (TPT) levels.
- We adapt SaCo [20] to rPPG regression as a quantitative faithfulness score.
- We evaluate UBFC-rPPG clip-level correlations between attribution metrics (skin coverage, SaCo) and prediction quality (waveform Pearson r, heart rate (HR) error).

2. Related Work

2.1. XAI Methods

Transformer explainability methods are commonly grouped into perturbation-, gradient-, and attention-based families [21, 22]. *Perturbation-based* methods such as LIME [23] and SHAP [24] modify inputs and observe prediction changes, whereas *gradient-based* methods such as saliency [25], LRP [26], and Grad-CAM [27] use gradient signals; Integrated Gradients [28] later appears as the F^c component of Beyond Intuition (Section 3.2). *Attention-based* methods visualize or aggregate attention weights: raw attention gives per-layer views, attention rollout [18] multiplies augmented attention matrices for transitive token relationships, attention flow [18] computes maximum flow on the attention graph, and Beyond Intuition [19] combines rollout with integrated gradients into attention perception ($P^{(L)}$) and reasoning feedback (F^c). Because RhythmFormer's spatial decisions occur in bi-level routing attention with top-k selection, we use four attention-based variants—raw, rollout, flow, and Beyond Intuition—spanning per-layer, cumulative, and gradient-augmented attribution to address the contested faithfulness of raw attention alone [15, 16, 17, 29].

2.2. XAI in rPPG

Existing rPPG XAI works—Dual-path TokenLearner [10], PhysKANNet [13], DD-rPPGNet [14], and TS-CAN+ [12]—use raw attention, gradient saliency, or Grad-CAM visualizations, with TS-CAN+ showing responses to background or other non-facial regions; our manual inspection of its published Grad-CAM visualization also shows responses on red clothing, but none provides quantitative faithfulness evaluation. CIN-rPPG [30] instead inspects module-internal channel-spatial interactive learning maps rather than applying post-hoc XAI.

2.3. Quantification in XAI—Faithfulness Evaluation

A common quantitative approach is cumulative perturbation, which progressively removes salient pixels but confounds individual pixel-group effects because each step builds on previous removals [20]. SaCo [20] avoids this by perturbing each group individually and computing pairwise consistency between salience ranking and prediction impact. To our knowledge, this is the first adaptation of SaCo to rPPG regression.

3. Method

3.1. RhythmFormer Overview

RhythmFormer [9] is an end-to-end transformer that maps an RGB facial video $X \in \mathbb{R}^{3\times T\times H\times W}$ to an rPPG signal $\hat{y} \in \mathbb{R}^T$. Its architecture (Figure 1) has four components: (1) a **Fusion Stem** integrating raw

frames with temporal frame differences to capture BVP color variations; (2) **Patch Embedding** into non-overlapping tokens; (3) a three-level **Hierarchical Temporal Periodic Transformer** (TPT), with two TPT blocks per level, a Local Context Enhancement (LCE) branch, and periodic sparse attention at temporal resolution $T/2^n$ for level $n \in \{1, 2, 3\}$ (80, 40, 20 frames for a 160-frame clip); and (4) a **Predictor Head** producing the 1D rPPG waveform.

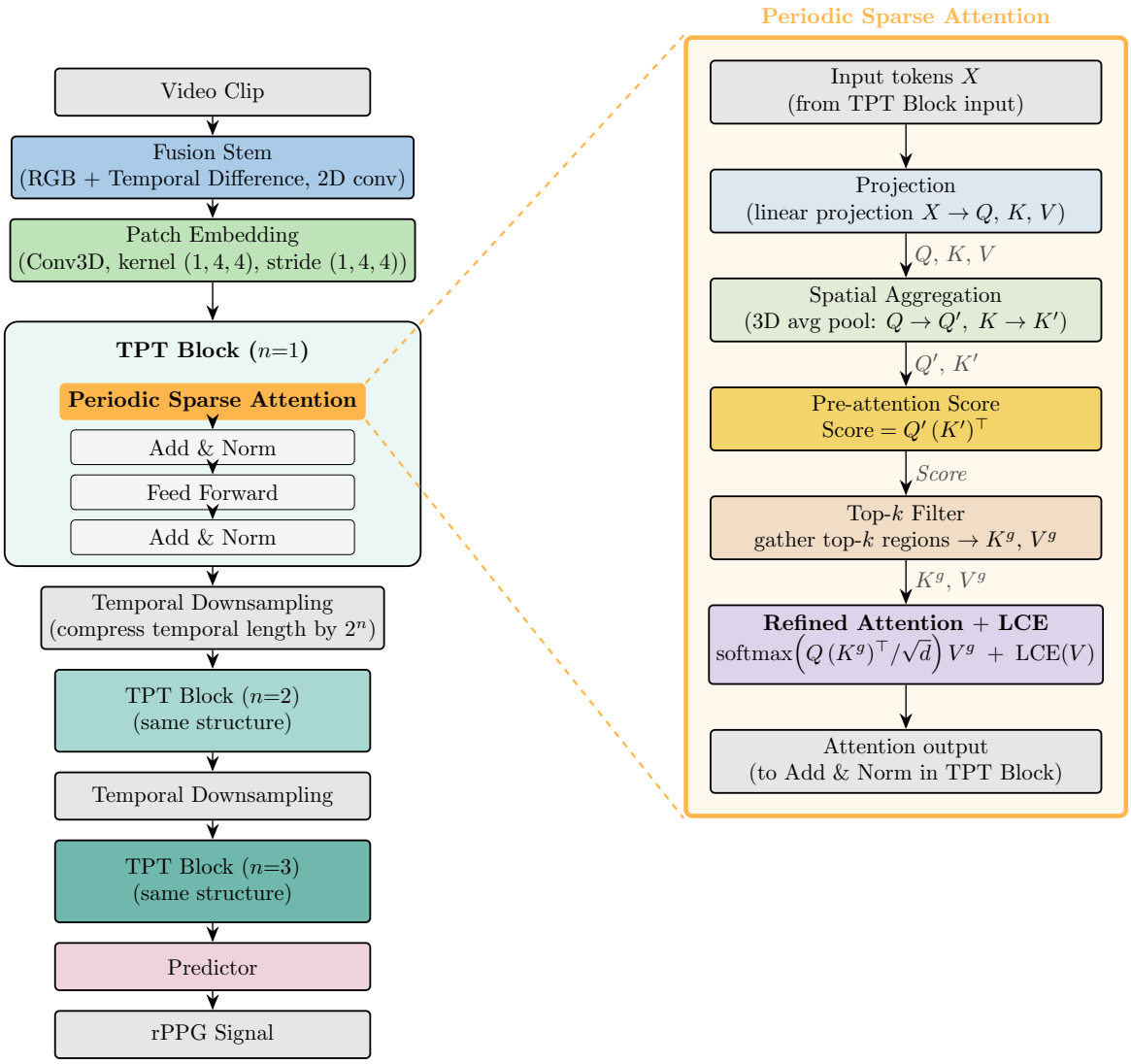


Figure 1: RhythmFormer architecture. *Left*: Fusion Stem, Patch Embedding, three TPT levels ($n = 1, 2, 3$) separated by Temporal Downsampling, and Predictor. *Right*: Periodic Sparse Attention: pre-attention aggregates Q, K and selects top-k regions K^g, V^g from $Q'(K')^\top$; refined attention applies scaled dot-product attention over gathered regions with LCE(V).

Periodic sparse attention. The key innovation is a two-stage attention mechanism in each TPT block (Figure 1). In *pre-attention*, queries Q and keys K are projected by temporal difference convolutions (TDC), average-pooled in 3D to Q' and K', scored as Score $= Q'(K')^\top$, and filtered by top-k selection. In *refined attention*, the gathered key-value pairs undergo softmax scaled dot-product attention:

$$SA = \text{Softmax}\left(\frac{Q(K^g)^\top}{\sqrt{d}}\right)V^g + \text{LCE}(V), \tag{1}$$

where K^g, V^g are the gathered key and value tokens selected by the pre-attention stage, and d denotes the block embedding dimension used by the released RhythmFormer/BiFormer implementation for

query–key scaling. In the reproduced checkpoint, this is the full embedding dimension (d=64), not the per-head dimension (16). We retain this released-code scale because changing it would define a different attention distribution from the trained model used for all reported experiments. The LCE branch, a depthwise 3×3×3 convolution on V, supplies an implicit positional inductive bias because RhythmFormer omits explicit positional embeddings. All six Bi-Level Routing Attention (BRA) layers use top-k routing with k=40. The candidate region set contains 640, 640, and 320 regions in the three TPT levels, respectively, so top-40 retains roughly 6.3%, 6.3%, and 12.5% of candidate regions per layer. These values follow the default RhythmFormer configuration [9].

Selection rationale and remaining gap. We select RhythmFormer because it reports competitive rPPG performance across benchmarks [9] while retaining a single-stream transformer design. This supports cumulative attention methods (rollout, flow, Beyond Intuition), which trace cross-layer attention chains; unlike architectures with parallel pathways or external modules such as PhysFormer++ [8], Dual-path TokenLearner [10], and CodePhys [31], RhythmFormer gives a clean sequential chain across TPT_1, TPT_2, and TPT_3.

3.2. Attention-Centered Attribution Methods

We extract attention matrices from all six attention layers (three TPT levels × two blocks) and adapt raw attention, attention rollout, attention flow, and Beyond Intuition. The two BRA blocks within TPT level s are denoted TPT_s-B0 and TPT_s-B1. Each TPT block yields *coarse pre-attention* scores at region granularity (4×4, N=16) and *refined attention* weights at token granularity (8×8, N=64) after top-k selection and softmax. Because refined attention is sparse, we reconstruct it as a dense $N\times N$ matrix with zeros at excluded positions for rollout, flow, and Beyond Intuition. Raw attention, rollout, and flow are applied to both pre-attention and refined matrices; Beyond Intuition is applied only to refined matrices. All matrices are head-averaged before aggregation, so $A^{(l)}$ denotes head-mean attention at layer l. For raw attention, rollout, and attention flow, the two temporal axes of each attention tensor are averaged at its native TPT resolution, leaving a purely spatial 16×16 pre-attention matrix or 64×64 refined-attention matrix before method-specific aggregation. Only Beyond Intuition retains the temporal axis across layers and unifies the three TPT resolutions to T=20 before its cross-layer product.

Raw attention. For each layer l, we compute received attention for spatial position k as $r_k^{(l)} = \sum_q A_{qk}^{(l)}$. This column-wise aggregation measures how much each position is attended to as a key, replacing the [CLS]-row aggregation used in ViT methods [32], which RhythmFormer lacks. Results are grouped by TPT level (averaging within each level) to yield three maps per resolution.

Attention rollout. Following [18], we recursively multiply augmented attention matrices $\hat{A}^{(l)} = 0.5\,I + 0.5\,A^{(l)}$, where the identity term models the residual skip connection with roughly equal weight: $\tilde{A} = \hat{A}^{(L)} \cdots \hat{A}^{(1)}$. Without a [CLS] token, we use column-sum aggregation, corresponding to input-side saliency under our top-down convention (rows: query/output; columns: key/input): $\psi_k = \sum_q \tilde{A}_{qk}$.

Attention flow. Following [18], we model each attention weight $A_{ts}^{(l)}$ as the capacity of an edge from token s at layer $l-1$ to token t at layer l, then compute maximum flow between all input–output token pairs. For each input token i at layer 0, we sum the max-flow values to every output token j at layer L: $\phi_i = \sum_j \text{MaxFlow}(v_i^{(0)}, v_j^{(L)})$. This gives the information capacity flowing out of input position i, complementing the column-sum saliency used for raw attention and rollout.

Beyond Intuition. Beyond Intuition [19] decomposes attribution as

$$T = P^{(L)} \odot F^{c}, \tag{2}$$

where $\odot$ denotes the Hadamard product. $P^{(L)}$ is the W-corrected attention rollout. Here, vanilla rollout is the baseline before BI's corrections: it multiplies attention matrices and implicitly treats attention weight as downstream influence, equivalent to assuming identity value projection W. Because each token is transformed into $z_j W$, a token with large attention weight but small $\|z_j W\|$ propagates little information. We therefore scale each token's rollout column by $\alpha_j = \|z_j W\|/\|z_j\|$, where $W = W_{\text{out}} W_v$, down-weighting tokens whose projected magnitude is small relative to their input magnitude. F^c is the integrated gradient of the loss w.r.t. the last-layer attention:

$$F^c = \text{ReLU}\left(-\frac{1}{N_{\text{steps}}} \sum_{s=1}^{N_{\text{steps}}} \nabla_{A^{(L)}} \mathscr{L}\left(\frac{s}{N_{\text{steps}}} \cdot X\right)\right), \tag{3}$$

where gradients are averaged over N_{steps} interpolation steps from a zero baseline to the input to mitigate saturation; because RhythmFormer minimizes training loss, the gradient is negated so ReLU retains positions whose increase reduces loss. The Hadamard product combines structural importance from rollout with decision-relevant gradient signal. Following the original formulation [19], which computes F^c from the last transformer layer's attention, we apply BI only to refined attention matrices (8×8), where spatial decisions are made, rather than pre-attention routing scores (4×4). We adapt BI to RhythmFormer with four modifications:

- *Sparse-to-dense reconstruction.* The W-correction preserves the top-k-excluded zeros since $\alpha_j \cdot 0 = 0$, so positions the model does not attend to remain zero throughout $P^{(L)}$.
- *Temporal resolution unification.* All layers are downsampled to T=20 to enable cross-level matrix multiplication in $P^{(L)}$; this unification is specific to Beyond Intuition.
- *Mean-intensity baseline.* Because inputs are z-score normalized per channel and clip, the effective baseline at k=0 is a uniform video at the per-channel per-clip mean, matching SaCo (Section 3.3), rather than an all-zero black video. This removes BVP-related variation while staying within the training distribution.
- *LCE branch excluded from W.* We fold only the linear chain $W = W_{\text{out}} W_v$ into the W-correction and leave LCE outside W. LCE is a depthwise 3×3×3 convolution, not a token-wise linear map, and therefore does not match BI's linear-projection derivation [19]; it also applies a level-fixed spatial kernel rather than content-routed value projection. A convolution-aware BI extension is left to future work.

3.3. Faithfulness Evaluation via SaCo

We adopt SaCo [20] for explanation faithfulness. Given a saliency map, the input is partitioned into K equal-sized groups ranked by salience, with $s(G) = \sum_{p \in G} M_p$ denoting the total saliency over group G. For each pair (G_i, G_j) with $s(G_i) \geq s(G_j)$, SaCo tests whether masking G_i has greater prediction impact than masking G_j. The faithfulness coefficient is:

$$F = \frac{\sum_{i<j} w_{ij}}{\sum_{i<j} |w_{ij}|} \in [-1, 1], \tag{4}$$

where $w_{ij} = +(s(G_i) - s(G_j))$ when masking G_i has greater impact than masking G_j, and $w_{ij} = -(s(G_i) - s(G_j))$ when the ordering is inverted. Because the signed weight is proportional to the saliency gap, larger gaps contribute more to F. Thus F=1 indicates perfect faithfulness, F= − 1 full inversion, and $|F|$ the correlation degree between saliency ranking and prediction impact [20].

Adaptation for regression. We replace the classification confidence drop with prediction distance:

$$\nabla\text{pred}(x, G_i) = \text{MAE}\Big(\Phi(x),\ \Phi(x_{G_i})\Big), \tag{5}$$

where $\Phi(x)$ is the predicted rPPG waveform and x_{G_i} replaces pixels in G_i with the per-clip per-channel mean intensity, following SaCo's Algorithm 1 to keep perturbations within the training distribution

while removing G_i's spatial content. A larger MAE indicates stronger model reliance on G_i. Using $\Phi(x)$ as the reference, rather than the ground-truth label, preserves the key property of SaCo [20]: impact is determined solely by model sensitivity, without conflating faithfulness with prediction accuracy.

3.4. Skin Attention Quantification

Because rPPG originates from blood-volume-induced color variation on exposed skin, off-skin attention cannot carry pulsatile information; we therefore hypothesize that higher skin coverage should be associated with lower prediction error. We quantify heatmap-skin overlap using a face parser built on the BiSeNet architecture [33]. To align grids, each raw attention heatmap (4×4 or 8×8; Section 3.2) is min-max normalized to $[0, 1]$ and bicubic-upsampled to $H{\times}W{=}128{\times}128$, yielding $\mathbf{h} \in [0, 1]^{H\times W}$. The binary skin mask $\mathbf{m} \in \{0, 1\}^{H\times W}$ encodes each pixel as skin (1) or non-skin (0). All skin-coverage analyses use face-parser classes $\{1, 2, 3, 10, 14\}$ (facial skin, both eyebrows, nose, and neck). This extended mask is the canonical definition for the principal comparison, artifact-ablation sensitivity analysis, and Beyond Intuition component decomposition. The skin coverage metric is:

$$\text{Skin Coverage} = \frac{\sum_i \mathbf{h}_i \cdot \mathbf{m}_i}{\sum_i \mathbf{h}_i}, \tag{6}$$

where i indexes all $H{\times}W$ pixels. The ratio reports the fraction of total attention mass falling on skin. A value of 1.0 indicates all attention falls on skin; 0.0 indicates no overlap.

Coverage, concentration, and the role of SaCo. Eq. 6 is a normalized mass ratio and is therefore sensitive to attribution concentration as well as skin alignment: concentrating mass on a few high-value pixels can increase coverage. This matters because Beyond Intuition is structurally sparser than attention-only methods: $T = P^{(L)} \odot F^c$ multiplies value-projection-weighted rollout by a gradient-supported mask, zeroing part of the heatmap before normalization (Section 3.2). Thus higher skin coverage reflects both spatial alignment and concentration, which the metric cannot separate. We therefore do not treat skin coverage as standalone faithfulness; SaCo (Section 3.3) provides the independent check by perturbing top-ranked regions and measuring prediction impact.

3.5. Waveform-Level Signal Quality

To relate attention patterns to signal-level prediction quality, we measure per-clip Pearson correlation between predicted and ground-truth rPPG waveforms:

$$r = \frac{\sum_{t=1}^{T}(\hat{y}_t - \overline{\hat{y}})(y_t - \overline{y})}{\sqrt{\sum_{t=1}^{T}(\hat{y}_t - \overline{\hat{y}})^2 \cdot \sum_{t=1}^{T}(y_t - \overline{y})^2}}, \tag{7}$$

where $\hat{y}$ and y are the predicted and ground-truth rPPG waveforms within a single clip of T frames. This captures waveform reconstruction fidelity and differs from the HR-level Pearson ρ in Table 1, which measures agreement of scalar heart-rate estimates across subjects.

3.6. Heart-Rate-Level Error

For clip-level HR analysis in the XAI correlation figures, we estimate scalar heart rate from predicted and ground-truth rPPG waveforms using RhythmFormer's FFT-based protocol [9]. Each XAI point is one UBFC-rPPG clip with T=160 frames at f_s=30 Hz, a 5.33 s window. Both waveforms are detrended (λ=100), bandpass filtered to 0.75–2.5 Hz (45–150 bpm), and converted to power spectral density by Welch's method. The heart rate is the dominant spectral peak in this band:

$$\text{HR}(y) = 60 \underset{f\in[0.75,\,2.5]}{\arg\max}\ \text{PSD}_y(f). \tag{8}$$

The clip-level HR error is

$$e_{\mathrm{HR}} = |\mathrm{HR}(\hat{y}) - \mathrm{HR}(y)|\,, \tag{9}$$

The quantity used for each clip-level correlation point is this absolute HR error. When aggregating over a set of clips, HR MAE is the mean of e_{HR} across clips.

4. Experiments

4.1. Experimental Setup

Scope. This evaluation is confined to UBFC-rPPG and to the reproduced RhythmFormer checkpoint (Section 4.2). This scope is intentional: the paper establishes an XAI framework for periodic sparse attention, rather than a cross-dataset or cross-model benchmark. Accordingly, conclusions about the generality of multi-hop leakage, skin-coverage rankings, and SaCo behaviour require validation on additional datasets (PURE [34], COHFACE, MMPD [35], V4V) and sparse-attention rPPG transformers (PhysFormer [8], EfficientPhys [5], and RhythmFormer variants with different top-k settings), which we leave to future work.

Dataset. We evaluate on UBFC-rPPG [36], which contains 42 videos at 640 × 480 resolution and 30 fps, each approximately one minute long (~1,800 frames), recording subjects in a stationary state. Following the original RhythmFormer protocol [9], the first 30 samples are used for training and the remaining 12 for testing. Each test video is segmented into non-overlapping 160-frame clips (`chunk_num` = ⌊frame_count/160⌋); the trailing segment shorter than 160 frames is discarded, yielding n=141 clips across the 12 test subjects (12 clips per subject for 11 of the 12 subjects; one subject with a shorter recording yields 9). We focus on UBFC-rPPG for two reasons. First, this work is a method paper aimed at establishing an XAI framework rather than benchmarking across datasets, and UBFC's controlled recording conditions reduce dataset-side confounders (motion artefacts, illumination shifts, parsing failures) so that method-level effects such as multi-hop leakage are not entangled with dataset noise. Second, UBFC-rPPG is the dataset on which the original RhythmFormer paper [9] reports results, allowing a direct reproduction check against the published numbers (Sec. 4.2) before any XAI analysis is performed.

Training. The model is trained for 30 epochs with batch size 4, learning rate 9×10^{-3}, and clip length 160 frames on 4 NVIDIA RTX A5000 GPUs. The hybrid loss is defined as:

$$\mathscr{L} = 0.2 \cdot \mathscr{L}_{\mathrm{time}} + \mathscr{L}_{\mathrm{CE}} + \mathscr{L}_{\mathrm{HR}}, \tag{10}$$

where $\mathscr{L}_{\mathrm{time}} = 1 - \rho(\hat{y}, y)$ is the negative Pearson correlation, $\mathscr{L}_{\mathrm{CE}}$ is the cross-entropy over the predicted power spectrum and ground-truth heart-rate bin, and $\mathscr{L}_{\mathrm{HR}}$ is the heart-rate distribution loss used in the original RhythmFormer objective [9]. Data augmentation includes random temporal resampling conditioned on heart rate and random horizontal flips. No modifications are made to the original RhythmFormer architecture or hyperparameters.

XAI configuration. Attention matrices are extracted from all six BRA layers (three TPT levels × two blocks). Beyond Intuition uses N_{steps}=20 integration steps, following the lower bound recommended by Sundararajan et al. [28] for approximating the integral within 5%. SaCo evaluation uses K=8 spatial groupings with the MAE impact metric; this is a practical trade-off given that the 8×8 refined-attention grid bounds the maximum number of spatial groupings at 64, and larger K increases computational cost with diminishing evaluation granularity.

Rollout reconstruction analysis. To characterize how top-k masking interacts with cumulative rollout, we evaluated the refined 64×64 attention matrices on all n=141 test clips. For each matrix, the analysis uses the 4032 off-diagonal entries (64^2-64) and excludes the diagonal because it is preserved by the $\frac{1}{2}I$ residual term in the augmented rollout matrices. For each off-diagonal token pair, we record whether it is zeroed by top-k selection in at least one refined-attention layer and compare its final rollout value against an always-retained baseline consisting of entries that remain non-zero in all six refined-attention layers.

4.2. Training Reproduction

Table 1 compares our reproduced training results against the metrics reported in the original paper [9]. The close agreement confirms that our reproduction is faithful and the model checkpoint is reliable for subsequent XAI analysis. The slight discrepancies are expected: non-deterministic CUDA reductions under `DistributedDataParallel` cause the optimization trajectory to diverge across hardware configurations [37], and because the original protocol uses no independent validation set—selecting the best checkpoint directly on the 12 test subjects via the HR bias term B_{HR} [9]—this run-to-run variability propagates directly into the reported metrics.

Table 1
Comparison of our reproduced results with the original paper on UBFC-rPPG. ↓ = lower is better, ↑ = higher is better. An em dash indicates a value not reported by the cited source [9].

	MAE↓	RMSE↓	MAPE↓	ρ↑
Paper [9]	0.50	0.78	—	0.99
Ours	0.45	0.76	0.46	0.997

As an additional clip-level summary, we estimated heart rate and absolute error separately for each of the 141 non-overlapping 160-frame clips and then averaged the errors within and across subjects, yielding an MAE of 0.96 ± 0.16 beats per minute. This value is higher than the subject-level reproduction result of 0.45 beats per minute because the original protocol concatenates each subject's clip predictions before estimating a single heart rate, thereby providing greater spectral averaging. The clip-level summary does not replace any value used in the subsequent XAI analyses and does not change their results or interpretation.

5. Results

5.1. Qualitative Visualization

We first present qualitative visualizations to provide intuition for how different XAI methods interpret the model's attention. Figure 2 shows attention heatmaps averaged across all 12 test subjects and three TPT levels, revealing several patterns. First, all methods consistently highlight the facial skin region rather than background or non-skin areas. Second, comparing the two columns of attention-based methods reveals the effect of top-k selection: for raw attention, the refined column is visibly more concentrated on skin than the pre-attention column, directly reflecting the sparsification introduced by bi-level routing. However, this concentration difference is less pronounced for the cumulative methods. For rollout, the cumulative matrix product can restore attribution to positions excluded by individual layers' top-k; attention flow has the same path-based caveat because it aggregates capacity over the same multi-layer attention graph. We quantify the resulting coverage pattern below and discuss the rollout mechanism and its implication for flow in Section 6. Third, Beyond Intuition provides a complementary perspective: its final attribution $T = P^{(L)} \odot F^c$ is notably more concentrated than the attention-based methods, reflecting the sparsity inherited from top-k selection (Section 6).

Rollout and Beyond Intuition concentrate saliency on the cheeks and forehead—regions reported as strong rPPG signal ROIs [11]—whereas raw attention and attention flow spread more broadly. This qualitative ordering is quantified in the skin coverage analysis below.

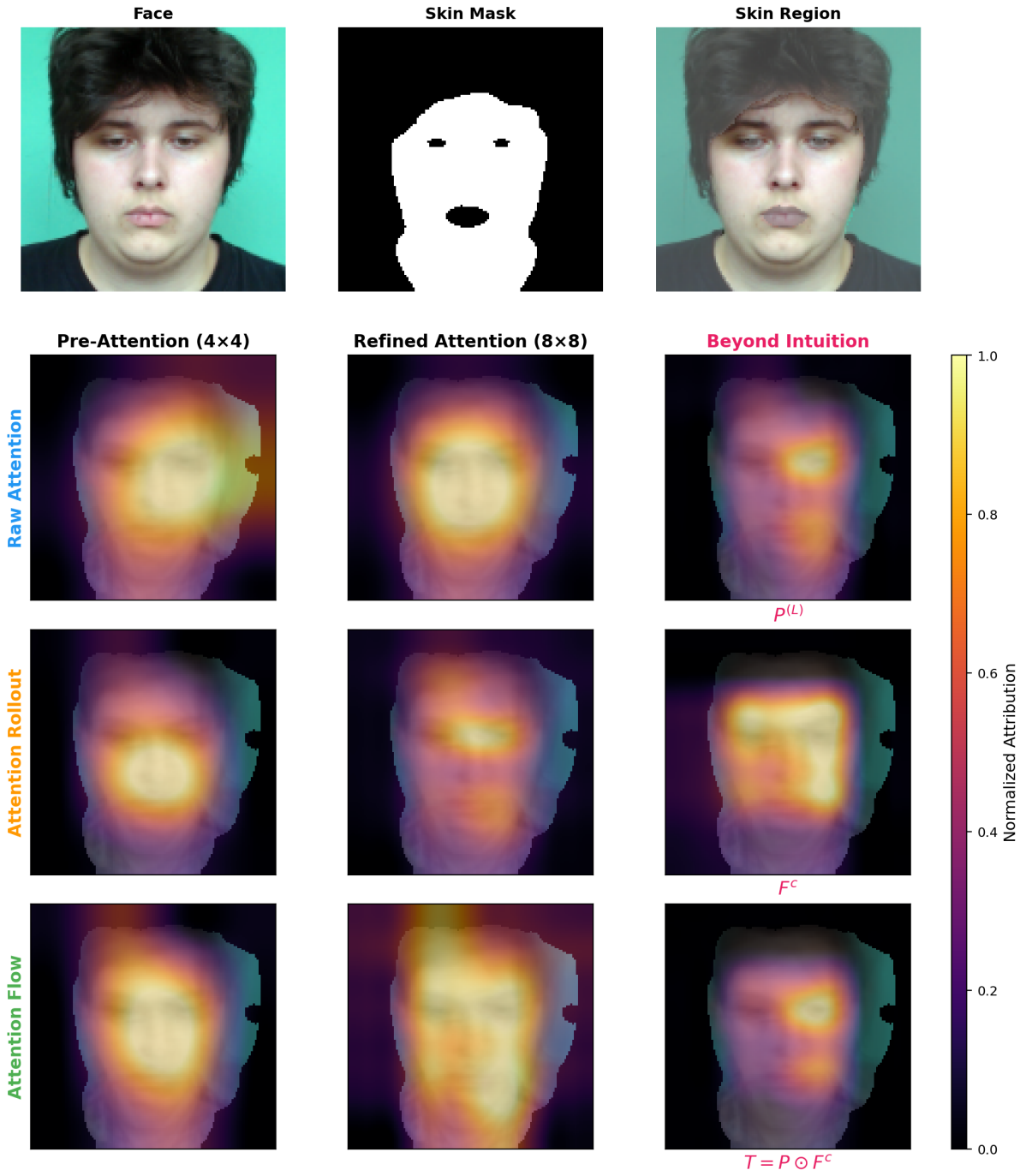


Figure 2: Pre-attention vs. refined attention with Beyond Intuition on UBFC-rPPG. Top row: representative sample face, skin mask, and skin region overlay. Middle 3 × 2 grid: rows = raw attention, rollout, and attention flow; columns = pre-attention (4×4) routing scores and refined (8×8) post-top-k weights, cross-subject and cross-TPT averaged. Right column: Beyond Intuition components $P^{(L)}$, F^c, and $T = P^{(L)} \odot F^c$, cross-subject averaged. Heatmaps are per-cell min-max normalized.

5.2. Skin Coverage Analysis

To move beyond individual examples, we quantify the overlap between attention heatmaps and physiologically relevant skin regions. Using the binary skin mask defined in Section 3.4, we compute the skin coverage ratio defined in Eq. 6. Since each BRA layer in RhythmFormer produces two independent attention matrices—pre-attention weights ($Q'K'^{\top}$, region-level) and refined attention weights (post-top-k softmax, token-level)—we compute skin coverage separately at each resolution: pre-attention at 4×4 and refined attention at 8×8. For cumulative methods (rollout, attention flow), each version operates on

the corresponding set of matrices independently.

Figure 3 compares the skin coverage distributions across all four methods and all test subjects. For each attention-based method (raw attention, rollout, attention flow), each data point represents the mean skin coverage of a single clip, averaged across the three TPT levels. Three patterns emerge. First, for raw attention, the refined stage achieves significantly higher skin coverage than the pre-attention stage ($\Delta = +0.22$, $p < 0.001$). Second, for rollout and attention flow, this relationship reverses: the refined stage yields significantly *lower* mean skin coverage than the pre-attention stage (mean refined-minus-pre differences $\Delta_{\text{mean}} = -0.12$ and -0.17, both $p < 0.001$). Third, Beyond Intuition achieves significantly higher skin coverage than all attention-based methods at the refined level ($p < 0.001$), but with markedly greater inter-subject variability—its distribution spans from below 0.4 to above 0.8. We analyze the mechanisms underlying these three phenomena in Section 6. Table 2 reports the median skin coverage alongside median SaCo faithfulness for each method.

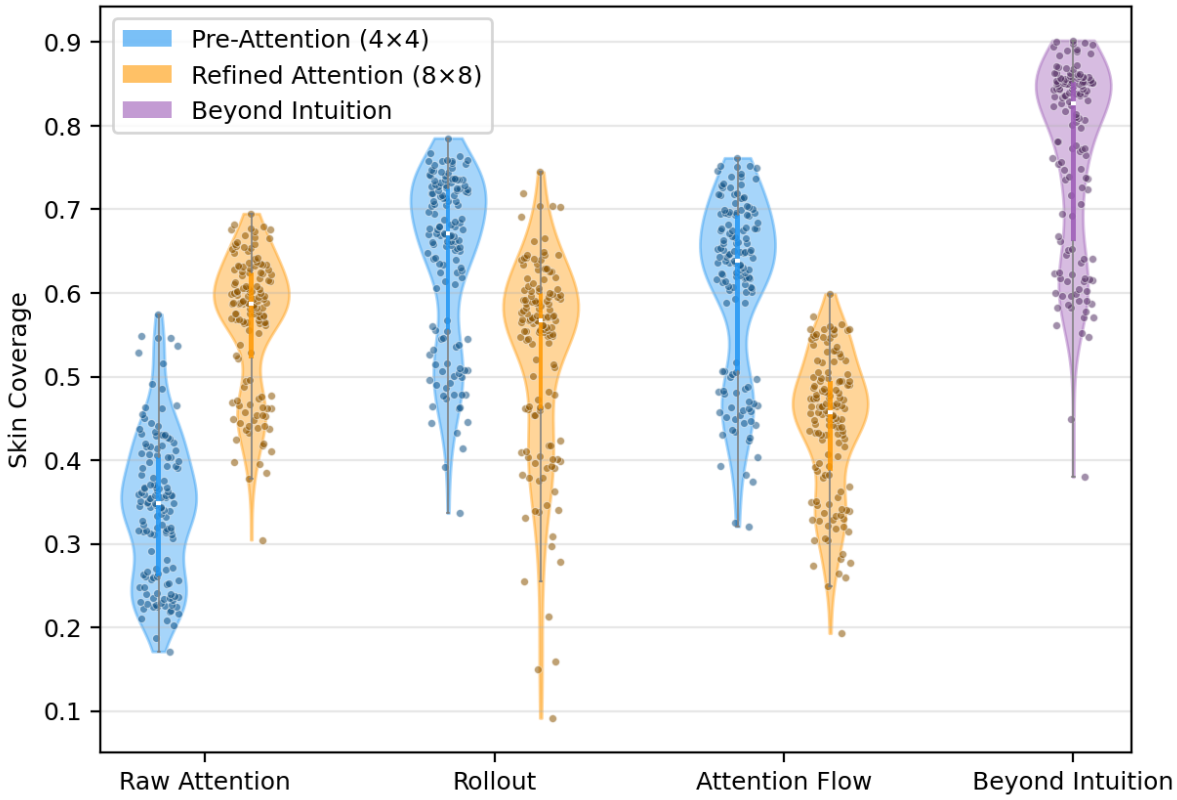


Figure 3: Violin plots of skin coverage for all four XAI methods. Pre-attention (4×4) and refined attention (8×8) are shown separately for attention-based methods. Beyond Intuition appears only at 8×8 because it operates on the refined attention matrices (Section 3.2): the pre-attention routing scores lack softmax normalization, making their gradient signal nearly uniform and uninformative for attribution.

Table 2
Median skin coverage and SaCo faithfulness (F, K=8) on UBFC-rPPG. For attention-based skin coverage, each clip is first averaged across the three TPT levels, and the table reports the median across clips. Values are reported to three decimal places for a uniform resolution across metrics. ↑ = higher is better. Best in **bold**. Beyond Intuition operates only at the refined level (Section 3.2).

	Skin Cov. (pre)↑	Skin Cov. (refined)↑	SaCo F↑
Raw Attention	0.348	0.587	0.846
Rollout	**0.671**	0.567	0.851
Attention Flow	0.638	0.457	0.846
Beyond Intuition	—	**0.826**	**0.917**

5.3. Rollout Reconstruction of Top-k-Masked Connections

Top-k selection explicitly zeros many refined-attention entries in individual layers. For example, TPT_1-B0 zeros $2036/4032 = 50.5\%$ of off-diagonal entries, and the six refined-attention layers contain 4197 such zero-entries on average over the 141 clips. The six-layer rollout restores every one of these excluded entries to a non-zero value: $R^{(L)}$ is fully dense, so 100% of the top-k-masked off-diagonal entries are recovered through multi-hop paths. This effect is not limited to the full six-layer traversal. Within a single TPT level, 98–100% of the first block's top-k zeros become non-zero after the two-layer rollout of that level. The recovered values are also not negligible: the median rollout value at recovered entries is 0.0106, only 1.8× smaller than at always-retained entries (0.0192), and the 99th percentile of recovered values (0.0392) exceeds the always-retained median. Overall, recovered entries account for approximately 48% of plain rollout's total off-diagonal mass.

5.4. Faithfulness Evaluation

Figure 4 shows the distribution of SaCo faithfulness scores across all test clips (median values in Table 2). Beyond Intuition achieves the highest overall faithfulness, but notably exhibits several clips with negative SaCo scores; other methods also contain low-scoring outliers at the subject level. We examine these outlier cases in Section 6.

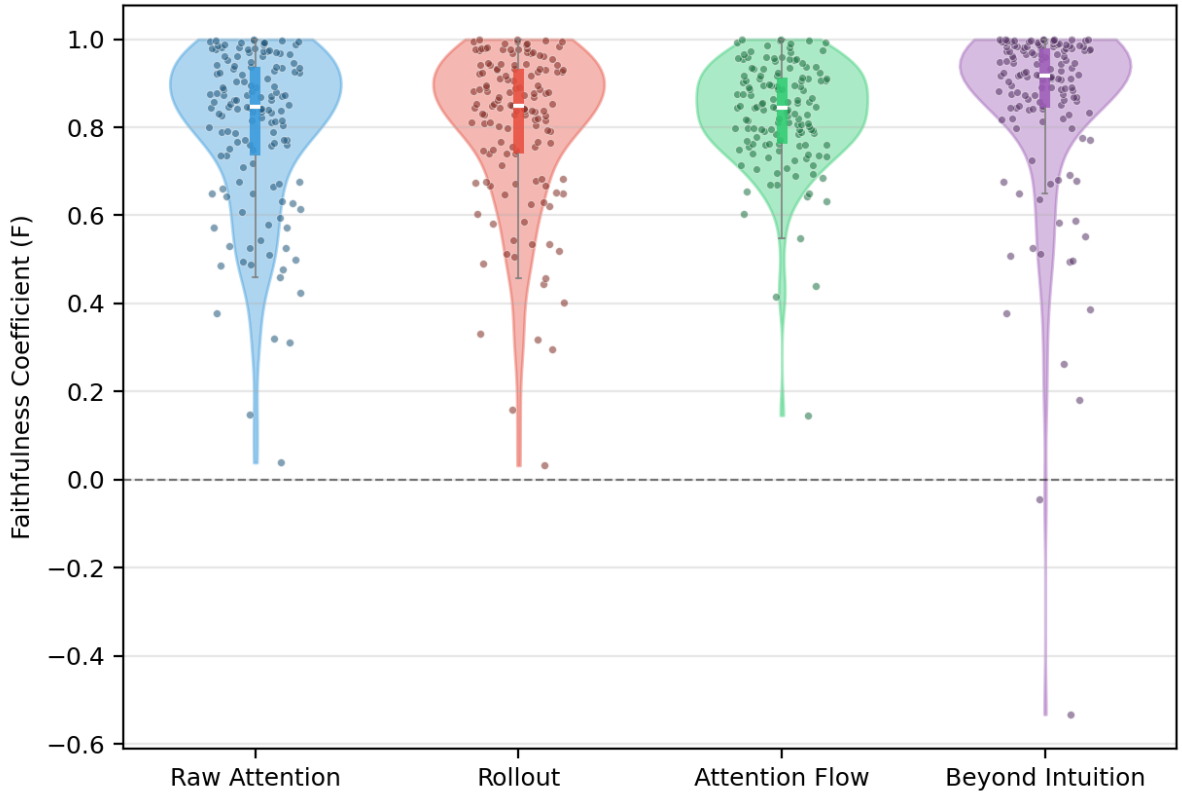


Figure 4: Violin plots of SaCo faithfulness scores by method across all test subjects. Beyond Intuition achieves the highest median faithfulness; the distributional shapes are broadly comparable across methods.

5.5. Clip-Level Correlation Analysis

The preceding sections evaluate XAI metrics in aggregate; here we examine clip-level relationships between two prediction-quality metrics—HR-level mean absolute error (MAE; FFT-based) and waveform-level Pearson r (time-domain)—and two XAI metrics—skin coverage and SaCo faithfulness—using Spearman's rank correlation coefficient. Each point represents a single 160-frame clip; points are colored by subject to visualize within- versus between-subject variability.

MAE vs. skin coverage. If skin coverage reflects the model attending to physiologically informative regions, we expect a negative correlation between skin coverage and MAE. Figure 6 matches this expected direction across all four methods, although the correlations are weak: attention rollout and attention flow are nominally below .05, while Beyond Intuition is nominally just above it. This indicates that HR accuracy is shaped by factors beyond spatial attention to skin.

MAE vs. SaCo faithfulness. We expect lower MAE to coincide with higher SaCo, i.e., clips on which the model predicts well should also be those on which its attribution is more faithful to its decision. Figure 5 does not exhibit this expected trend: no method shows a significant correlation. We discuss the underlying reasons in Section 6.

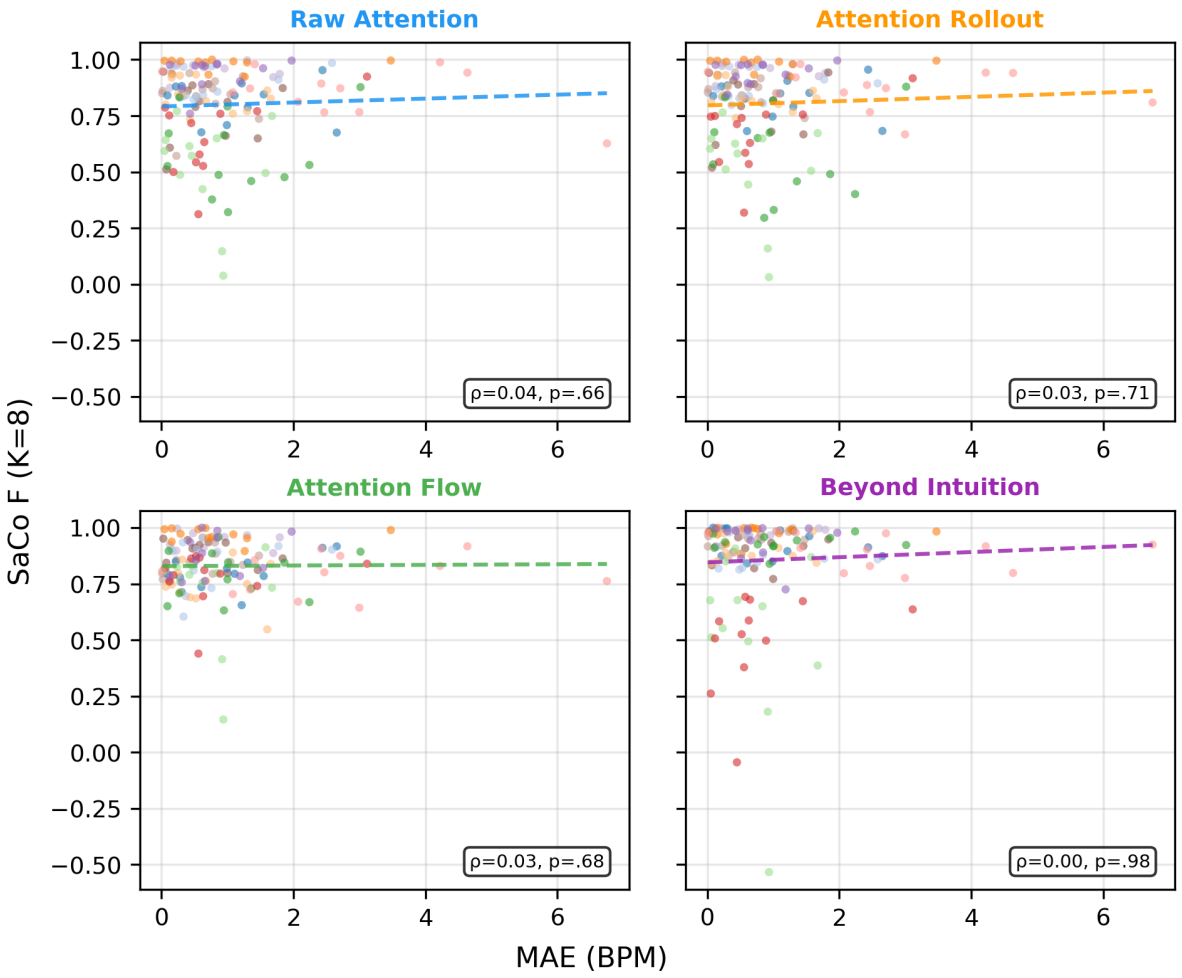


Figure 5: MAE vs. SaCo faithfulness (F, K=8) per clip for each XAI method (n=141 clips across 12 test subjects). Each point represents one 160-frame clip; colors denote different test subjects. Spearman correlations satisfy $|\rho| \leq 0.04$ and all nominal $p > .05$ (Beyond Intuition ρ=0.00, p=0.98). The 141 clips are nested within 12 subjects (encoded by colour); the reported Spearman ρ and p are clip-level summary statistics, not tests of independent observations.

Waveform quality vs. skin coverage. To disentangle waveform reconstruction quality from heart rate accuracy, we replace the MAE axis with the per-clip Pearson r (Eq. 7; Figure 7); we expect higher skin coverage to coincide with higher Pearson r, i.e., better waveform fidelity. Figure 7 partially matches this expectation: attention rollout and Beyond Intuition show significant positive correlations, while attention flow and raw attention do not reach significance. Notably, these positive correlations are stronger than the corresponding MAE–coverage correlations in Figure 6, indicating that skin attention tracks waveform fidelity more closely than HR accuracy.

Asymmetric impact of skin attention: waveform vs. HR. Both clip-level findings are consistent with the expectation that skin attention contributes to prediction quality: skin coverage shows a significant positive correlation with waveform fidelity (Figure 7; rollout ρ=0.31, Beyond Intuition ρ=0.24, both $p < 0.005$), and a directionally negative trend with HR error across all four methods (Figure 6). These two observations correspond to successive stages of the same propagation chain: spatial attention to skin yields cleaner BVP waveforms, which in turn yield sharper spectral peaks and lower HR error. The first stage (skin → waveform) is directly visible in our data; the second stage (waveform → HR) is attenuated under UBFC-rPPG's controlled recording conditions, where sufficiently high SNR allows even imperfect waveforms to produce unambiguous spectral peaks, saturating HR accuracy (MAE ≈ 0 for most subjects) and masking the influence of spatial attention at the HR level. Although this leaves only a weak negative MAE–skin correlation in the present data, we would expect this correlation to strengthen substantially in noisier datasets (motion artifacts, variable illumination, lower SNR), where the spectral margin between the HR peak and noise floor narrows and waveform-quality differences propagate more directly to HR error. Under this framing, skin coverage remains a meaningful XAI metric: it captures where the model directs spatial attention and correlates with prediction quality in the expected directions, with HR-level effects masked here only by UBFC's saturated conditions.

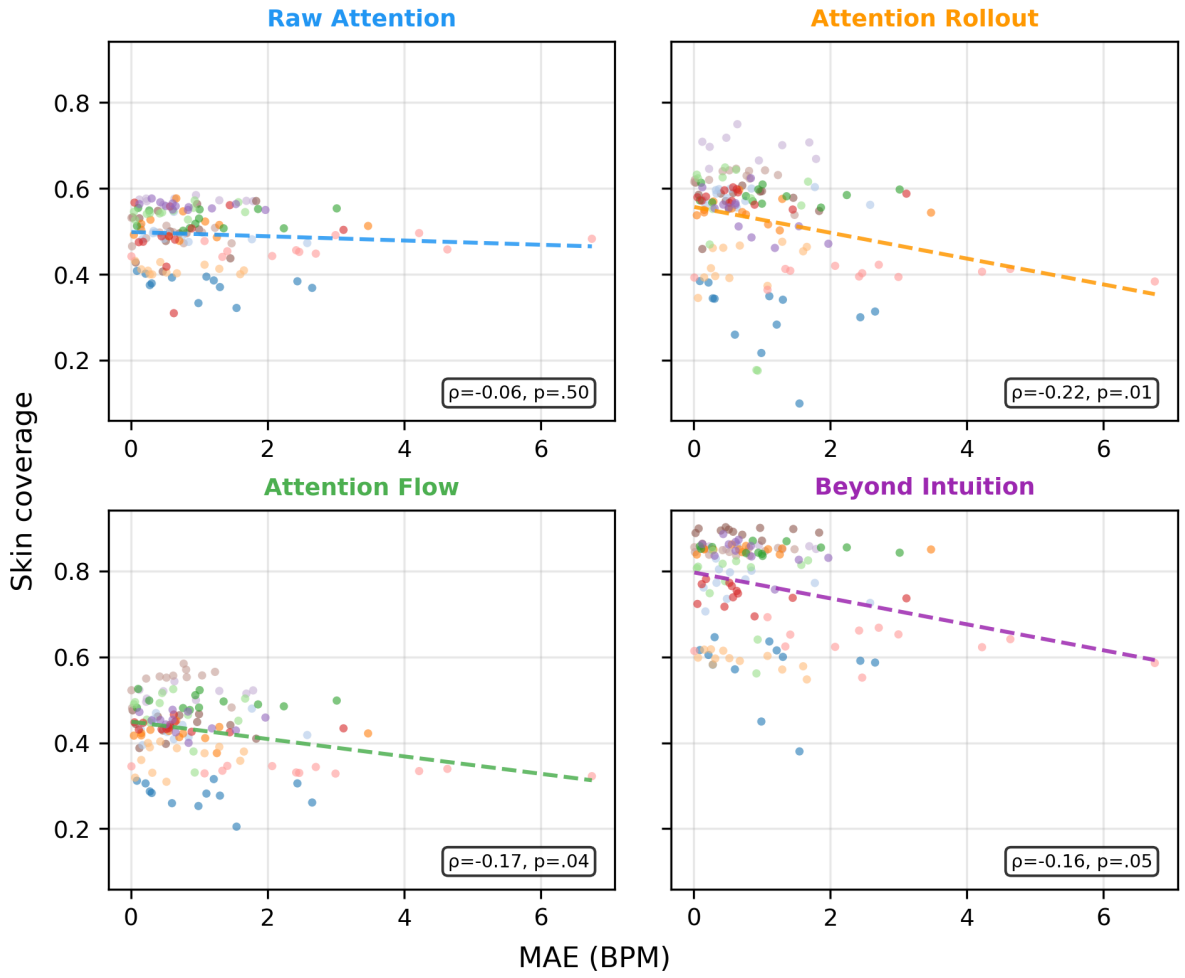


Figure 6: MAE vs. skin coverage per clip for each XAI method (n=141 clips across 12 test subjects). Each point represents one 160-frame clip; colors denote different test subjects. Spearman correlations: rollout ρ=−0.22 (p=0.01), flow ρ=−0.17 (p=0.04), Beyond Intuition ρ=−0.16 (p=0.05), raw ρ=−0.06 (p=0.50). The 141 clips are nested within 12 subjects (encoded by colour), so these clip-level ρ and p are descriptive summary statistics rather than tests of independent observations.

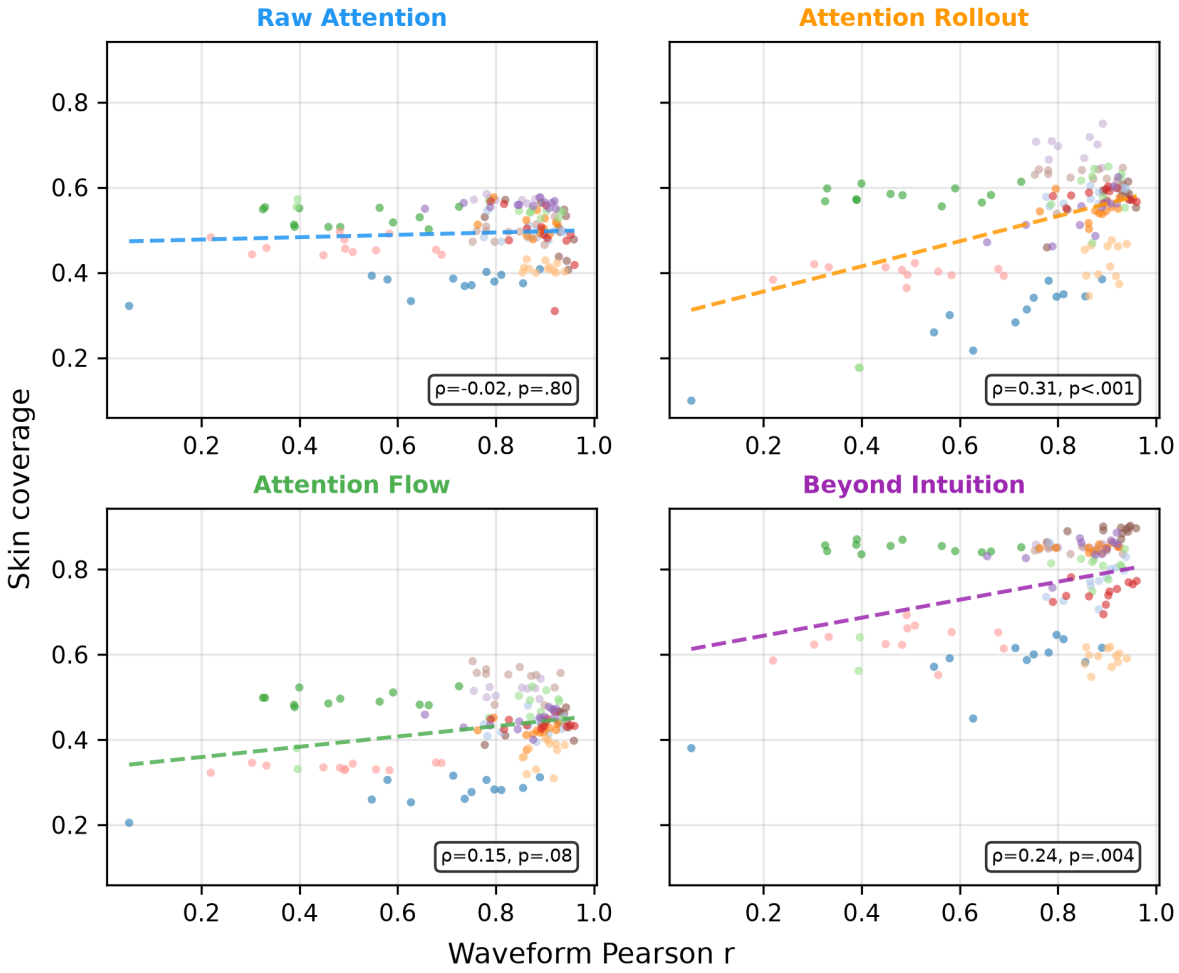


Figure 7: Waveform quality (Pearson r) vs. skin coverage per clip for each XAI method (n=141 clips across 12 test subjects). Higher Pearson r indicates better waveform reconstruction. Colors denote different test subjects. Spearman correlations: rollout ρ=0.31 ($p < 0.001$), Beyond Intuition ρ=0.24 (p=0.004), flow ρ=0.15 (p=0.08), raw ρ=−0.02 (p=0.80). The 141 clips are nested within 12 subjects (encoded by colour); these clip-level ρ and p are descriptive summary statistics, not tests of independent observations.

5.6. Artifact Ablation Case Study

To complement the aggregate analyses, we report one low-SaCo outlier as a result-level case study; the interpretation is deferred to Section 6. Subject 46 clip 11 has the lowest Beyond Intuition SaCo in the test set ($F = -0.53$), and frame-by-frame inspection of the raw video showed that the subject rubs their eye for 6 out of 160 frames (3.75% of the input). We replaced these six hand-occlusion frames with stable neighbouring frames from the same clip—preserving the 160-frame input length and the temporal context—and re-ran the full forward pass and XAI pipeline, so that only those 6 frames change. Figure 8 compares the original and ablated heatmaps, and Table 3 reports SaCo F and skin coverage before and after the intervention. Across all four attribution methods, both SaCo and skin coverage recover after the intervention; the per-method values are reported in Table 3.

Table 3
Artefact ablation on subject 46 clip 11. SaCo faithfulness F and skin coverage before and after replacing the 6 hand-artefact frames (3.75% of the clip) with stable neighbouring frames from the same clip. Skin coverage uses the canonical extended face-parser mask $\{1, 2, 3, 10, 14\}$ computed from original frame 0 for both conditions; raw, flow, and BI are reported as their per-clip average, and rollout uses the refined average. All four methods' SaCo recover toward $F \approx 0.8$ (Beyond Intuition shows the largest gain), while skin coverage moves in the same direction with smaller magnitudes. Values are reported to two decimal places.

	SaCo F			Skin coverage		
Method	Orig	Mod	Δ	Orig	Mod	Δ
Raw attention	0.04	0.84	+0.80	0.61	0.67	+0.06
Rollout	0.03	0.80	+0.77	0.12	0.44	+0.31
Attention flow	0.15	0.86	+0.72	0.25	0.39	+0.14
Beyond Intuition	−0.53	0.76	+1.29	0.73	0.80	+0.07

6. Discussion

Mechanism: how bi-level routing shapes attribution

Section 5.3 shows that plain rollout reconstructs top-k-masked refined-attention connections into non-zero multi-hop weights, with recovered entries contributing approximately 48% of rollout's off-diagonal mass. This multi-hop leakage explains the skin-coverage reversal observed for rollout in Section 5.2: even though top-k at every layer removes a large fraction of weight from non-skin positions, the rollout product carries roughly half of the off-diagonal mass back to excluded positions, so refined rollout coverage drops *below* pre-attention (Δ=−0.12, Fig. 3). Attention flow shows the same empirical refined-vs-pre drop (Δ=−0.17), and it is mechanistically exposed to the same sparse-routing caveat because max-flow is computed over paths in the same layered attention graph; however, we do not quantify its leakage mass here and therefore do not claim that its magnitude matches rollout. Visually the rollout heatmap still looks face-centred—the *peak* values remain on skin, dominated by always-retained attention—but the spread of recovered mass to non-skin pixels is large enough to drag the skin-coverage ratio down. Importantly, this leakage hurts *spatial* interpretability but not *faithfulness*: rollout and attention flow remain comparable to raw attention on SaCo (0.851 and 0.846 vs. 0.846, Tab. 2), indicating that their cumulative/path-based attributions still fall on regions the model genuinely relies on, even when those regions extend beyond the skin mask.

Beyond Intuition (BI) [19] forms its attribution as $T = P^{(L)} \odot F^c$, where $P^{(L)}$ is the W-corrected rollout (reweighting each token by its value-projection magnitude $\alpha_j = \|z_j W\| / \|z_j\|$) and F^c is a last-layer integrated-gradient mask. We compute skin coverage on every test clip (n=141) using the canonical extended mask $\{1, 2, 3, 10, 14\}$ and report the median across all clips for each component of the BI pipeline (Tab. 4). Replacing vanilla rollout with the W-corrected rollout $P^{(L)}$ raises median skin coverage from 0.57 to 0.68 (Δ=+0.11; 141/141 **clips improved**). Because every clip improves with no exception, the data directly show that the W-correction step attenuates the multi-hop leakage characterized above. The integrated-gradient mask F^c does not admit the same kind of direct comparison: it is constructed from a last-layer gradient rather than from a chain of sparse attention matrices, so there is no "uncorrected F^c" to serve as a fair baseline. (F^c on its own has median skin coverage 0.66, but this value cannot be interpreted as the effect of an isolated correction step.) Its effect is instead observed in combination with $P^{(L)}$: the Hadamard product $T = P^{(L)} \odot F^c$ raises median skin coverage further from 0.68 to 0.83 (Δ=+0.15; 141/141 **clips improved**).

Taken together, the two steps lift median skin coverage by +0.26 over vanilla rollout, with every clip improving at every step. These counts register the sign of each clip's change, not its size: six clips improve by under 0.02 in the first step, and a repeated face parse moves one below zero, while the medians shift by at most 0.003. Beyond Intuition is therefore not just nominally more faithful—it explicitly closes the multi-hop leakage gap that plain rollout suffers from, indicating that suppressing this leakage is the better attribution strategy under top-k sparse attention.

Table 4
BI decomposition under the canonical extended mask $\{1, 2, 3, 10, 14\}$ for n=141 clips across 12 test subjects. Improved reports clips with a positive pairwise difference.

Comparison	Δ median	Improved
$P^{(L)}$ vs vanilla rollout	+0.11	141/141
T vs $P^{(L)}$	+0.15	141/141
T vs F^c	+0.17	137/141

These results also argue against treating skin coverage as a standalone XAI metric for rPPG. Because the coverage ratio is sensitive to attribution concentration as well as to spatial alignment with skin (Section 3.4), and Beyond Intuition is structurally sparser than the attention-based methods, a single coverage number—whether the unweighted ratio or a planned region-weighted variant that up-weights forehead and cheek regions associated with stronger rPPG signal quality [11]—cannot by itself separate "attributes to the right regions" from "attributes more concentratedly." We therefore read skin coverage only in conjunction with the perturbation-based SaCo coefficient, which measures whether the highlighted regions are ones the model genuinely relies on and does not reward concentration as such.

Spatial attention does not uniformly track prediction quality

SaCo faithfulness shows no significant correlation with either HR error (Figure 5) or waveform quality, indicating that attribution consistency under perturbation captures an aspect of model behavior independent of prediction accuracy at both the waveform and heart rate levels. This independence does not diminish SaCo's utility: it evaluates whether the attribution map correctly identifies which spatial regions the model relies on, as verified by perturbation, regardless of whether the model's prediction is accurate. A clip can be poorly predicted yet faithfully attributed—the model's attention is genuinely concentrated on those regions, they simply do not carry a strong rPPG signal—or accurately predicted with low faithfulness, where the correct prediction arises from distributed attention that no single region dominates.

XAI variability under saturated performance. Even among accurately predicted clips, both skin coverage and SaCo faithfulness vary substantially rather than converging to uniformly high values. This indicates that strong predictive performance does not imply a uniform internal strategy: two clips with equally accurate heart rate predictions may rely on different spatial attention patterns—a distinction invisible to aggregate performance metrics.

Outlier analysis: artifact ablation

The artefact ablation reported in Section 5.6 (Fig. 8 and Table 3) is a single-clip empirical observation; here we discuss what it implies for SaCo as a cross-method metric. The intervention was triggered by Beyond Intuition's strongly negative score on this clip ($F = -0.53$): a negative F means the regions BI flags as most salient receive *lower* prediction impact than the regions it deems unimportant—attention anti-correlated with what actually drives the output—which made the clip a candidate for manual inspection rather than dismissal as random noise. Once the six hand-artefact frames are replaced, all four attribution methods recover toward $F \approx 0.8$, and because the same direction of change emerges across raw attention, rollout, attention flow, and Beyond Intuition—four algorithmically distinct attributions—a method-specific artefact explanation is unlikely; the four methods agree both that this clip is problematic and on the direction of change after the intervention, suggesting that SaCo behaves consistently across attribution families on this clip. Skin coverage moves in the same direction but is less responsive on this clip, consistent with the broader observation that skin coverage and SaCo capture related but non-identical aspects of attribution quality. We treat this as illustrative rather than confirmatory: a single-clip ablation cannot establish that SaCo measures genuine attention/impact alignment in general, and a systematic multi-outlier study would be required to support that stronger claim.

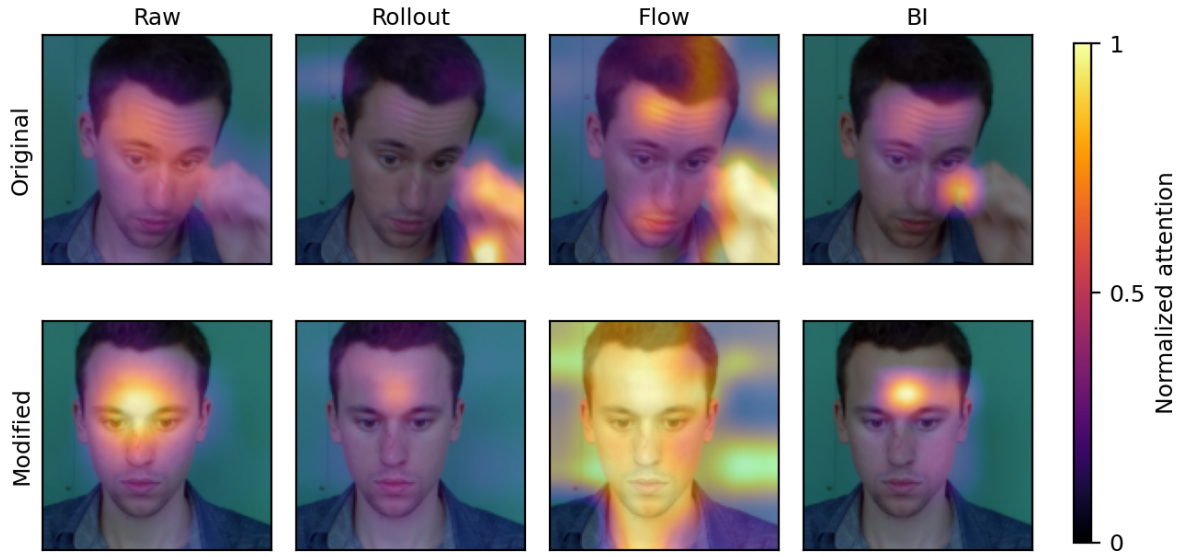


Figure 8: Hand occlusion ablation on subject 46 clip 11. Top: original; bottom: with the hand-occlusion frames replaced by stable neighbouring frames from the same clip. Across all four attribution methods (Raw, Rollout, Flow, BI), the replacement recovers a more skin-aligned distribution; BI SaCo lifts from −0.53 to 0.76 on this clip.

Limitations. Two caveats apply. **(i)** Our evaluation uses UBFC-rPPG, which has only 12 test subjects under controlled conditions (stable indoor lighting, near-frontal pose, no motion artefacts on most clips). RhythmFormer's HR error saturates near zero on these clips, so the dataset is not complex enough to test whether spatial attention metrics (skin coverage, SaCo) actually correlate with prediction quality in the regimes where the model is challenged. The weak MAE–skin correlation we observe could equally reflect an absence of relationship or a saturation effect that masks one—the present data cannot distinguish these. Moreover, the n=141 clips are nested within these 12 subjects (roughly twelve clips each), so they are not independent observations; the effective sample size at the subject level is small, and all reported p-values—including the borderline clip-level cases in Section 5.5—should be read as descriptive rather than confirmatory. **(ii)** Our analysis covers only UBFC-rPPG and only RhythmFormer. Whether the multi-hop leakage pattern reported here is specific to RhythmFormer's top-k schedule, or generalises to other sparse-attention rPPG models (e.g., PhysFormer variants, BiFormer-based pipelines on other tasks), remains an open question; patterns observed here should not be promoted to general claims about sparse vision transformers without cross-model replication.

Future directions. Several extensions follow naturally. First, replication on more challenging benchmarks (e.g., PURE [34], COHFACE, MMPD [35], V4V), where motion artefacts and variable illumination push the model out of UBFC saturation, would let us test whether skin coverage and SaCo track prediction quality once HR error is non-trivial, and on additional sparse-attention rPPG transformers (e.g., PhysFormer [8], EfficientPhys [5], and RhythmFormer variants with different top-k settings) would test whether the multi-hop leakage pattern is specific to RhythmFormer's top-k schedule or generalises. The four attribution adaptations, the skin coverage metric together with a planned perfusion-weighted extension, and the regression-adapted SaCo coefficient introduced here are directly reusable for that follow-up evaluation. Near-term follow-up analyses should also compute attribution entropy from the refined attention tensors, complete the subject-median correlation analysis using per-clip prediction errors, and evaluate perfusion-weighted skin coverage from attention tensors and face parses. Second, our integrated-gradient baseline for F^c is currently the per-clip mean-intensity reference; alternative baselines (e.g., black, Gaussian noise, or learned) may shift F^c and consequently T, and a systematic comparison would clarify how baseline choice affects the attribution. Third, RhythmFormer's three TPT levels process features at different spatial and temporal scales; computing BI attribution per TPT level rather than only at the last layer could reveal whether each level emphasises distinct physiological

features—for example, whether earlier levels capture coarse skin-region attention and later levels refine to specific pulsatile regions.

7. Conclusion

This work compares four XAI attribution methods for transformer-based rPPG—raw attention, attention rollout, attention flow, and Beyond Intuition—using skin coverage over canonical face regions and SaCo faithfulness. Applied to RhythmFormer on UBFC-rPPG, skin coverage shows the expected moderate positive correlation with waveform-level Pearson and weak negative trend with HR MAE, consistent with skin-directed attention contributing to prediction quality, although the modest strengths may reflect UBFC's saturated HR regime or a genuinely weak relationship. SaCo is not strongly correlated with per-clip MAE, but in one low-SaCo outlier (subject 46 clip 11), replacing the offending frames improves SaCo in the same direction across all four methods; this suggests cross-method consistency on that clip, while a single case cannot rule out method-specific biases.

These findings challenge the dense-attention assumption behind cumulative attention attribution under sparse routing. For attention rollout [18], RhythmFormer's top-k bi-level scheme [38] lets multi-hop paths re-fill entries that individual layers zero out—multi-hop leakage—so filtered positions still accumulate roughly half of rollout's off-diagonal attribution mass. Attention flow [18] warrants the same caution because it also aggregates over the multi-layer attention graph via max-flow, although we do not quantify its leakage separately. Beyond Intuition reduces this rollout-specific issue: value-projection weighting attenuates leaked mass, and the integrated-gradient mask restricts attribution to gradient-supported regions. We also extend Beyond Intuition from classification [19] to regression by using the MAE between original and perturbed rPPG waveforms as the SaCo [20] impact signal.

These observations suggest two directions. First, on UBFC-rPPG, Beyond Intuition achieves the highest median skin coverage and SaCo while remaining robust to multi-hop leakage, although with greater inter-clip variability than rollout and attention flow. It is therefore a strong candidate for sparse-routing rPPG attribution on this dataset and model, while rollout or flow may remain preferable when stable per-clip behaviour matters more than typical performance. Second, replication on benchmarks with motion artefacts and variable illumination [34, 35] and on other sparse-attention rPPG transformers [8, 5] is needed to test whether the MAE–skin-coverage relationship sharpens outside UBFC saturation and whether multi-hop leakage generalises beyond RhythmFormer. The attribution adaptations, skin coverage metric with a planned perfusion-weighted extension, and regression-adapted SaCo coefficient are reusable for that evaluation.

Acknowledgments

This work was supported by the Taiwan National Science and Technology Council (NSTC) under Grant 111-2221-E-006-186 and Grant 114-2221-E-006-089.

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Anthropic) in order to: grammar and spelling checking, code generation for the analysis pipeline and figure-plotting scripts, and LaTeX formatting assistance. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] W. Verkruysse, L. O. Svaasand, J. S. Nelson, Remote plethysmographic imaging using ambient light, Optics express 16 (2008) 21434–21445.

[2] G. de Haan, V. Jeanne, Robust pulse rate from chrominance-based rPPG, IEEE Transactions on Biomedical Engineering 60 (2013) 2878–2886.
[3] W. Wang, A. C. den Brinker, S. Stuijk, G. de Haan, Algorithmic principles of remote PPG, IEEE Transactions on Biomedical Engineering 64 (2017) 1479–1491.
[4] W. Chen, D. McDuff, Deepphys: Video-based physiological measurement using convolutional attention networks, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 356–373.
[5] X. Liu, B. Hill, Z. Jiang, S. Patel, D. McDuff, EfficientPhys: Enabling simple, fast and accurate camera-based cardiac measurement, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 4997–5006.
[6] Z. Yu, W. Peng, X. Li, X. Hong, G. Zhao, Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 151–160.
[7] Z. Yu, Y. Shen, J. Shi, H. Zhao, P. H. Torr, G. Zhao, PhysFormer: facial video-based physiological measurement with temporal difference transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4176–4186.
[8] Z. Yu, Y. Shen, J. Shi, H. Zhao, Y. Cui, J. Zhang, P. Torr, G. Zhao, PhysFormer++: Facial video-based physiological measurement with SlowFast temporal difference transformer, International Journal of Computer Vision 131 (2023) 1307–1330. All Open Access, Hybrid Gold Open Access.
[9] B. Zou, Z. Guo, J. Chen, J. Zhuo, W. Huang, H. Ma, Rhythmformer: Extracting patterned rPPG signals based on periodic sparse attention, Pattern Recognition 164 (2025) 111511.
[10] W. Qian, D. Guo, K. Li, X. Zhang, X. Tian, X. Yang, M. Wang, Dual-path tokenlearner for remote photoplethysmography-based physiological measurement with facial videos, IEEE Transactions on Computational Social Systems 11 (2024) 4465–4477.
[11] S. Kwon, J. Kim, D. Lee, K. Park, Roi analysis for remote photoplethysmography on facial video, in: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2015, pp. 4938–4941.
[12] Y. Li, J. Huang, J. Zhao, D. Wu, M. Zheng, Ts-can+: A improved ts-can architecture for non-contact heart rate measurement, IEEE Transactions on Consumer Electronics 71 (2025) 1393–1401.
[13] T. Liu, H. Xiao, Y. Sun, K. Zuo, Z. Li, Z. Yang, S. Liu, Physkannet: A kan-based model for multiscale feature extraction and contextual fusion in remote physiological measurement, Biomedical Signal Processing and Control 100 (2025) 107111.
[14] P.-K. Huang, T.-H. Chen, Y.-T. Chan, K.-W. Chen, C.-T. Hsu, DD-rPPGNet: De-interfering and descriptive feature learning for unsupervised rPPG estimation, IEEE Transactions on Information Forensics and Security 20 (2025) 4956–4970.
[15] S. Jain, B. C. Wallace, Attention is not explanation, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 3543–3556.
[16] S. Wiegreffe, Y. Pinter, Attention is not not explanation, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, 2019, pp. 11–20.
[17] S. Serrano, N. A. Smith, Is attention interpretable?, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2019, pp. 2931–2951.
[18] S. Abnar, W. Zuidema, Quantifying attention flow in transformers, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 4190–4197.
[19] J. Chen, X. Li, L. Yu, D. Dou, H. Xiong, Beyond intuition: Rethinking token attributions inside transformers, Transactions on Machine Learning Research (2023).
[20] J. Wu, W. Kang, H. Tang, Y. Hong, Y. Yan, On the faithfulness of vision transformer explanations,

in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10936–10945.

[21] M. Mersha, K. Lam, J. Wood, A. K. AlShami, J. Kalita, Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction, Neurocomputing 599 (2024) 128111.

[22] D. Bhati, F. Neha, M. Amiruzzaman, A survey on explainable artificial intelligence (xai) techniques for visualizing deep learning models in medical imaging, Journal of Imaging 10 (2024) 239.

[23] M. T. Ribeiro, S. Singh, C. 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, 2016, pp. 1135–1144.

[24] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems 30, 2017, pp. 4765–4774.

[25] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, in: 2nd International Conference on Learning Representations, ICLR 2014, Workshop Track Proceedings, 2014.

[26] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PloS one 10 (2015) e0130140.

[27] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, International Journal of Computer Vision 128 (2019) 336–359.

[28] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: Proceedings of the 34th International Conference on Machine Learning, volume 70 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 3319–3328.

[29] D. Pruthi, M. Gupta, B. Dhingra, G. Neubig, Z. C. Lipton, Learning to deceive with attention-based explanations, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4782–4793.

[30] Q. Li, D. Guo, W. Qian, X. Tian, X. Sun, H. Zhao, M. Wang, Channel-wise interactive learning for remote heart rate estimation from facial video, IEEE Transactions on Circuits and Systems for Video Technology 34 (2024) 4542–4555.

[31] S. Chu, M. Xia, M. Yuan, X. Liu, T. Seppanen, G. Zhao, J. Shi, Codephys: Robust video-based remote physiological measurement through latent codebook querying, IEEE Journal of Biomedical and Health Informatics 29 (2025) 4932–4945.

[32] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: 9th International Conference on Learning Representations, ICLR 2021, OpenReview.net, 2021.

[33] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, N. Sang, BiSeNet: Bilateral segmentation network for real-time semantic segmentation, in: Computer Vision – ECCV 2018, 2018, pp. 334–349.

[34] R. Stricker, S. Müller, H.-M. Groß, Non-contact video-based pulse rate measurement on a mobile service robot, in: The 23rd IEEE International Symposium on Robot and Human Interactive Communication, IEEE, IEEE, Edinburgh, Scotland, UK, 2014, pp. 1056–1062. PURE dataset.

[35] J. Tang, K. Chen, Y. Wang, Y. Shi, S. Patel, D. McDuff, X. Liu, Mmpd: multi-domain mobile video physiology dataset, in: 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE, 2023, pp. 1–5.

[36] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, J. Dubois, Unsupervised skin tissue segmentation for remote photoplethysmography, Pattern Recognition Letters 124 (2019) 82–90.

[37] PyTorch Contributors, Reproducibility, https://pytorch.org/docs/stable/notes/randomness.html, 2024. Accessed 2026-05-07.

[38] L. Zhu, X. Wang, Z. Ke, W. Zhang, R. W. H. Lau, BiFormer: Vision transformer with bi-level routing attention, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 10323–10333.