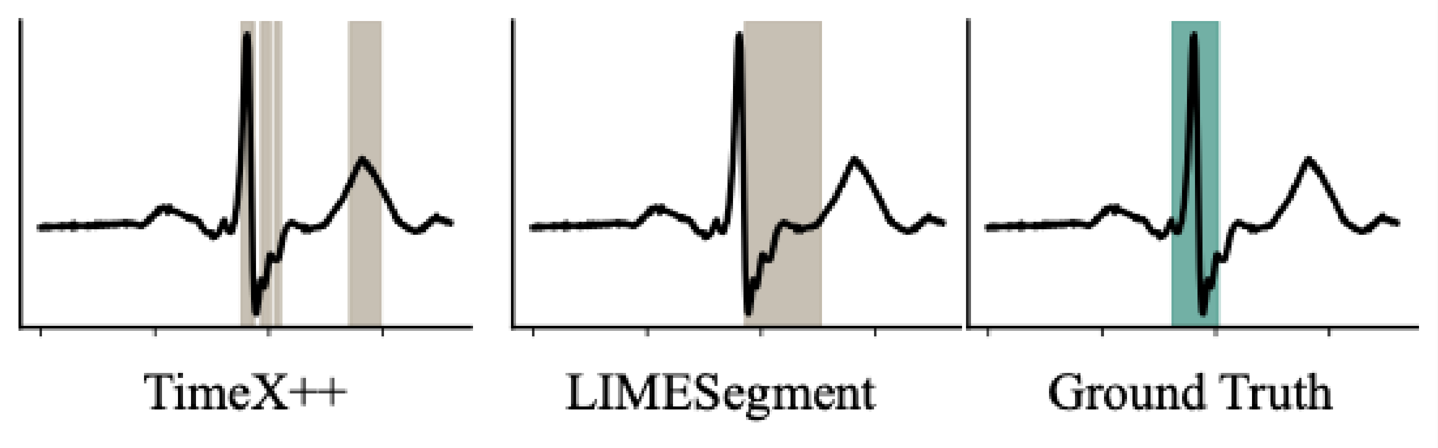
An Information-Theoretic Segment-wise Explainer for Time-Series Predictions

Hwijin Kim, Jaeho Kim, Changhee Lee

Motivations

1. Limitations & Motivations

- **Point-wise**
- **Patch-wise**



- It requires identifying both **the location of time series signals** that drive model predictions and **their matching to an interpretable temporal pattern**[1]
- We need explanations that are **temporally connected** and **visually digestible**[1]

[1] *O. Queen et al., "Encoding Time-Series Explanations through Self-Supervised Model Behavior Consistency," NeurIPS 2023.*

Motivations

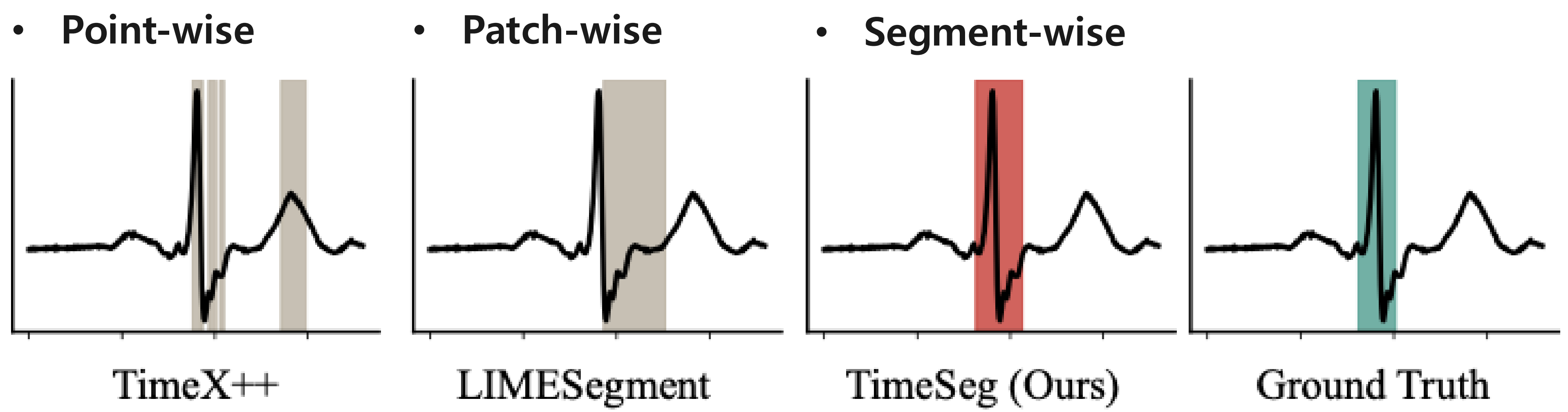
2. Related works

Approach	Method		Explanations	Localization	Black-box	Inference
Point-wise	Saliency Mask	Dynamask	✗ Scattered Time Point ✗ Fragmented Explanations	✓ Precise Localization	✓	Per-sample
		Extramask			✓	Per-sample
	Distribution Shift	FIT			✓	Per-sample
		WinIT			✓	Per-sample
	TimeX				✗	Amortized
	TimeX++				✗	Amortized
Patch-wise	LIMESegment		✓ Contiguous Explanations	✗ Fixed Length Patch ✗ Miss Boundary	✓	Per-sample
	SpectralX				✓	Per-sample
Segment-wise	TimeSeg		✓ Contiguous Explanations	✓ Variable Length Segments ✓ Precise Boundary	✓	Amortized

Motivations

2. Related works

Appr…					…erence
Poi…					…mple
					…mple
					…mple
					…mple
					…tized
					…tized
Patch-wise	LIMESegment	✓ Contiguous Explanations	✗ Fixed Length Patch ✗ Miss Boundary	✓	Per-sample
	SpectralX			✓	Per-sample
Segment-wise	TimeSeg	✓ Contiguous Explanations	✓ Variable Length Segments ✓ Precise Boundary	✓	Amortized



Problem Formulation

1. Notations

Univariate Time-series

$$\mathbf{x} = (x_1, \ldots, x_T) \in \mathbb{R}^T$$

Time-series Classifier

$$g_\theta : \mathbb{R}^T \to \mathbb{R}^C$$

Problem Formulation

1. Notations

Univariate Time-series

$$\mathbf{x} = (x_1, \ldots, x_T) \in \mathbb{R}^T$$

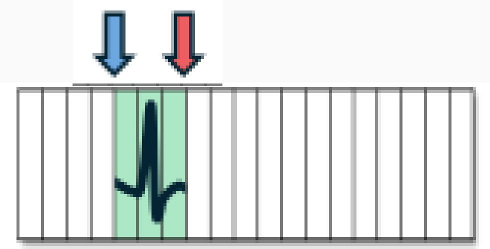
Black-Box g_θ

Time-series Classifier

$$g_\theta : \mathbb{R}^T \to \mathbb{R}^C$$

Segment Index

$$s = (t^s, t^e)$$



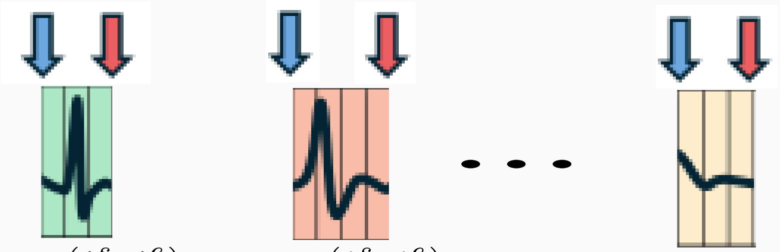
where, $t^s, t^e \in [T], t^s \leq t^e$

Segment (Contiguous Subsequence)

$$\mathbf{x}_s = \mathbf{x}_{t^s,t^e} := (x_{t^s}, \ldots, x_{t^e})$$

Segment Index set

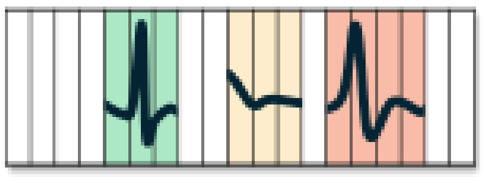
$$\mathbf{s}_{1:K} = \{s_1, s_2, \ldots, s_K\}$$



$s_1 = (t_1^s, t_1^e)$ $s_2 = (t_2^s, t_2^e)$ $s_K = (t_K^s, t_K^e)$

Segment set

$$\mathbf{x}_{\mathbf{s}_{1:K}} = \{\mathbf{x}_{s_1}, \mathbf{x}_{s_2}, \ldots, \mathbf{x}_{s_K}\}$$



Problem Formulation

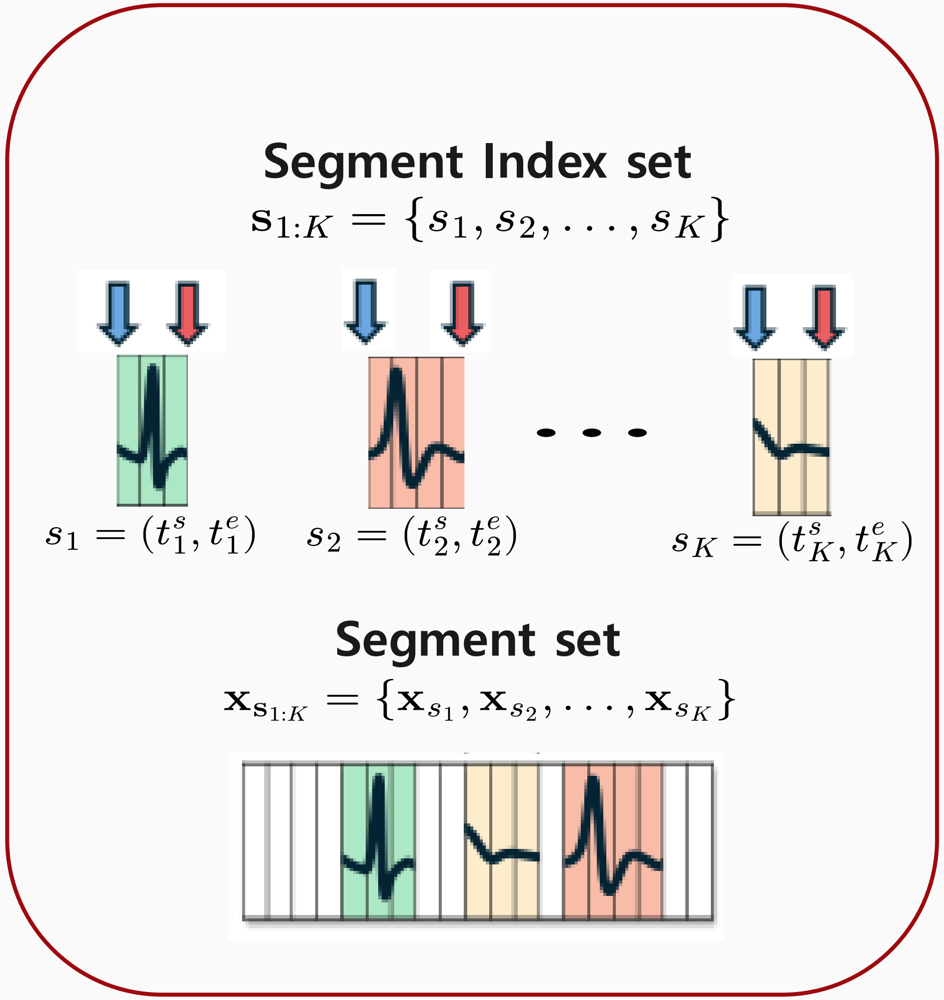
2. Segment-wise Explainer

Definition 1: Optimal Segment-wise Explainer

$$\mathcal{E}^* = \arg\max_{\mathcal{E}} I\Big(g_\theta(\mathbf{X}); \mathbf{X}_{\mathbf{s}_{1:K}}\Big) - \lambda J\Big(\mathbf{s}_{1:K}\Big)$$

$$\text{subject to} \quad \mathbf{s}_{1:K} \sim \mathcal{E}(\mathbf{X})$$

- $g_\theta(\mathbf{X})$: Output Prediction
- $I(\cdot,\cdot)$: Mutual Information
- $J(\cdot)$: Regularization for Complexity



Problem Formulation

2. Segment-wise Explainer

Definition 1: Optimal Segment-wise Explainer

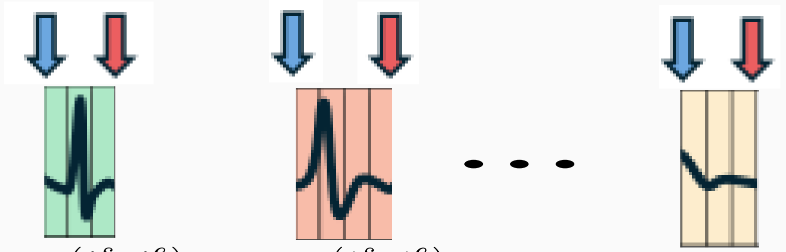
$$\mathcal{E}^* = \arg\max_{\mathcal{E}} I\Big(g_\theta(\mathbf{X}); \mathbf{X}_{\mathbf{s}_{1:K}}\Big) - \lambda J\Big(\mathbf{s}_{1:K}\Big)$$

$$\text{subject to} \quad \mathbf{s}_{1:K} \sim \mathcal{E}(\mathbf{X})$$

- $g_\theta(\mathbf{X})$: Output Prediction
- $I(\cdot,\cdot)$: Mutual Information
- $J(\cdot)$: Regularization for Complexity

Segment Index set

$\mathbf{s}_{1:K} = \{s_1, s_2, \ldots, s_K\}$



$s_1 = (t_1^s, t_1^e)$ $s_2 = (t_2^s, t_2^e)$ $s_K = (t_K^s, t_K^e)$

Segment set

Combinatorial Search

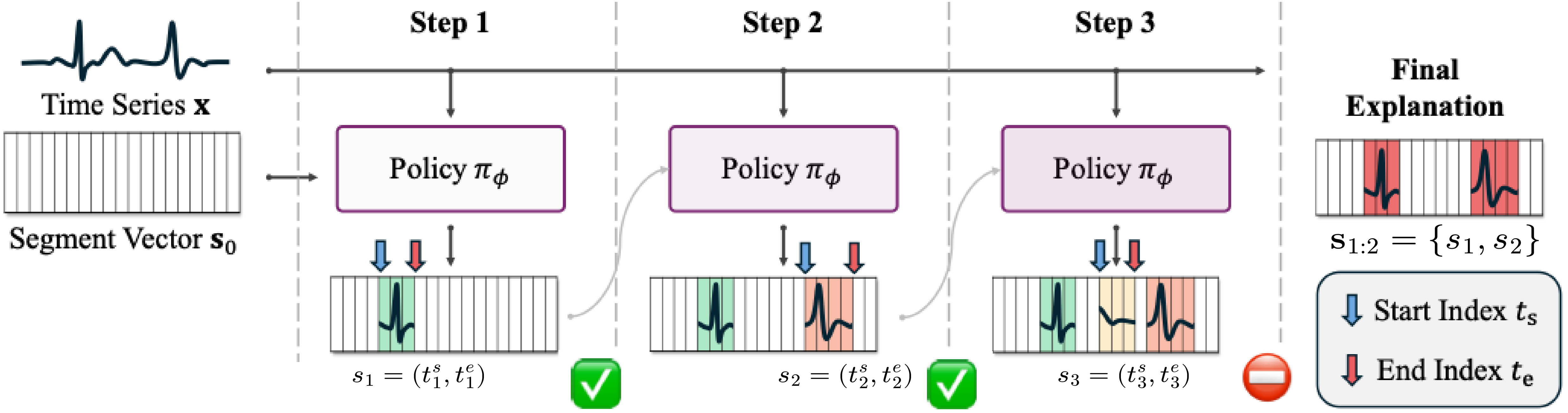
Exponential search space

K segments yield $\binom{T+1}{2K}$ candidate sets → $O(2^T)$

Method: TimeSeg

1. Reformulation as a Sequential Process

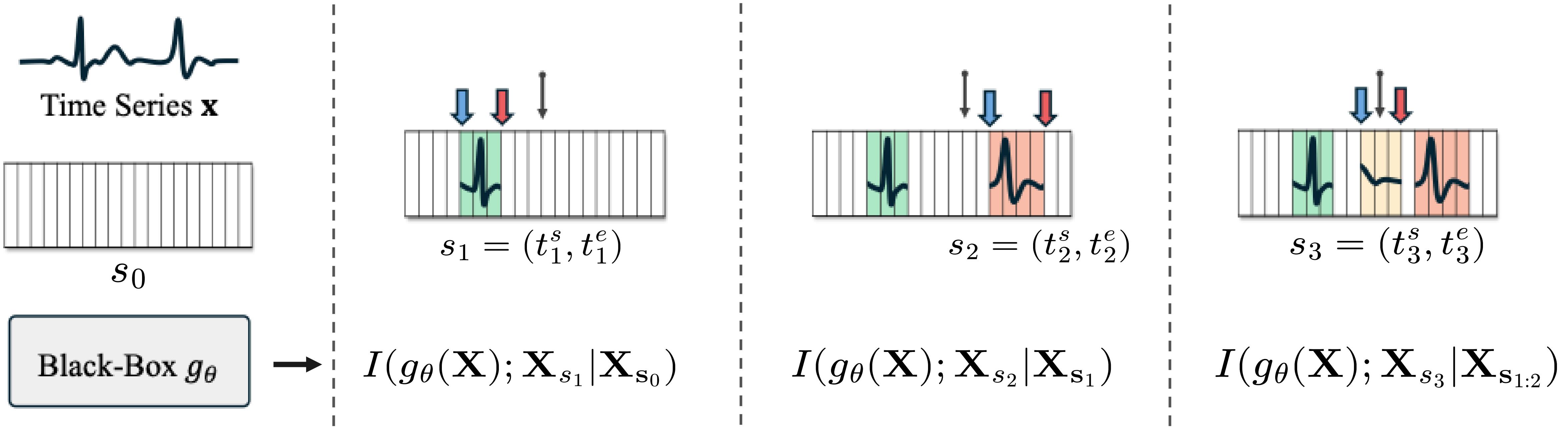
- Reformulate segment selection as a sequential process (MDP)
- Termination Criterion (Accept / Reject)
- Learn a stochastic policy via reinforcement learning



Method: TimeSeg

2. Chain Rule

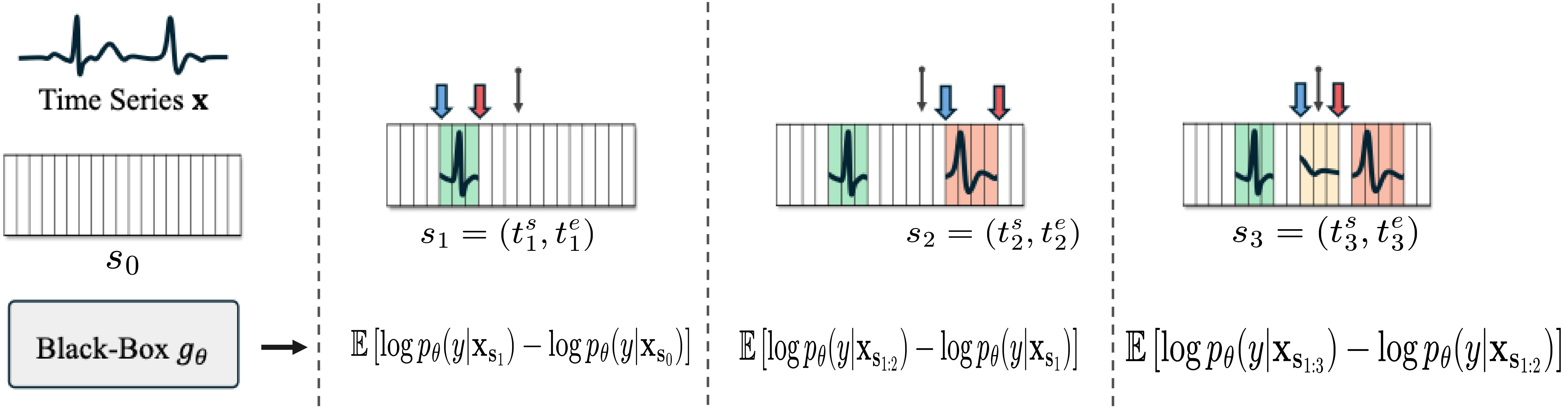
$$I\Big(g_\theta(\mathbf{X}); \mathbf{X}_{\mathbf{s}_{1:K}}\Big) = \sum_{k=1}^{K} I\Big(g_\theta(\mathbf{X}); \mathbf{X}_{s_k} \mid \mathbf{X}_{\mathbf{s}_{1:k-1}}\Big) \quad \textbf{(4.1)}$$



Method: TimeSeg

3. Difference between two Cross Entropy

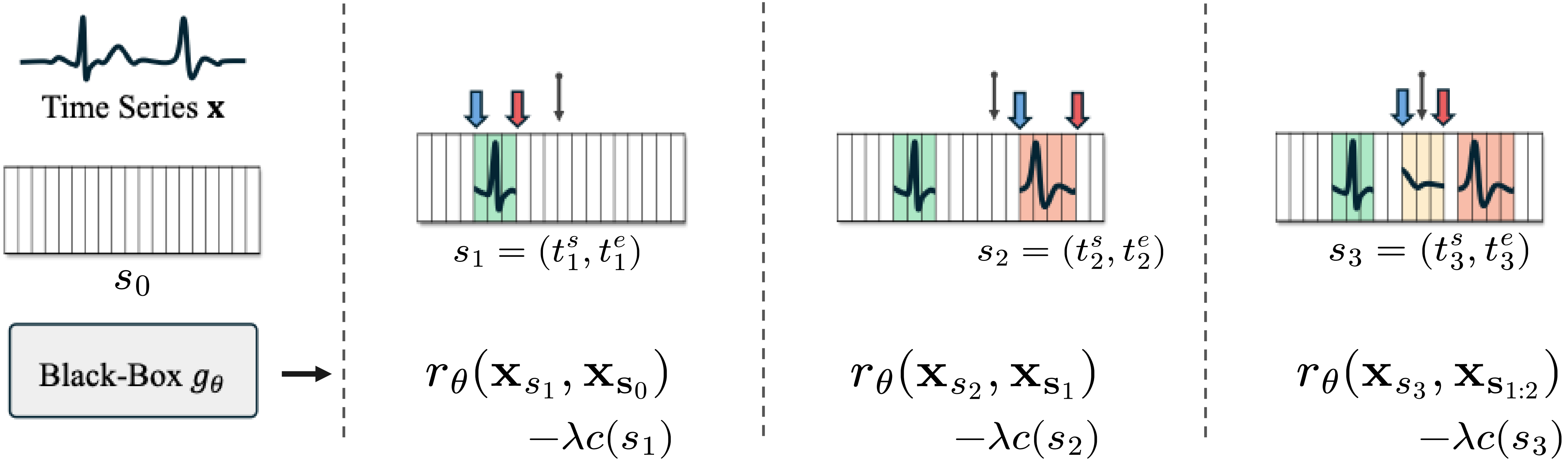
$$I(Y; \mathbf{X}_{s_k} \mid \mathbf{X}_{\mathbf{s}_{1:k-1}}) = \mathbb{E}_{\mathbf{x}_{\mathbf{s}_{1:k}}, y}\left[\log p_\theta(y \mid \mathbf{x}_{\mathbf{s}_{1:k}}) - \log p_\theta(y \mid \mathbf{x}_{\mathbf{s}_{1:k-1}})\right] \quad (4.3)$$



Method: TimeSeg

4. Reward + Penalty

$$r_\theta(\mathbf{x}_{s_k}, \mathbf{x}_{\mathbf{s}_{1:k-1}}) = \mathbb{E}_{p_\theta(y|\mathbf{x})}\Big[\log p_\theta(y|\mathbf{x}_{\mathbf{s}_{1:k}}) - \log p_\theta(y|\mathbf{x}_{\mathbf{s}_{1:k-1}})\Big] - \lambda c(s_k)$$



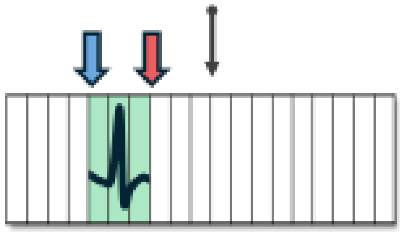
Method: TimeSeg

5. Termination Criterion

$$\frac{r_\theta(\mathbf{x}_{s_k}, \mathbf{x}_{\mathbf{s}_{1:k-1}})}{\mathbb{E}_{p_\theta(y|\mathbf{x})}\left[-\log p_\theta(y \mid \mathbf{x}_{\mathbf{s}_{1:k-1}})\right]} \leq \tau \rightarrow$$

Accept / Reject

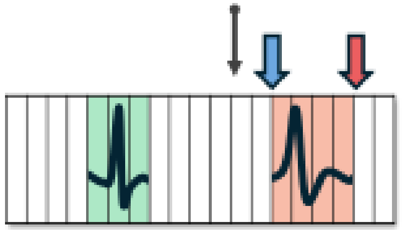
✅ / ⛔



$s_1 = (t_1^s, t_1^e)$

$r_\theta(\mathbf{x}_{s_1}, \mathbf{x}_{\mathbf{s}_0})$

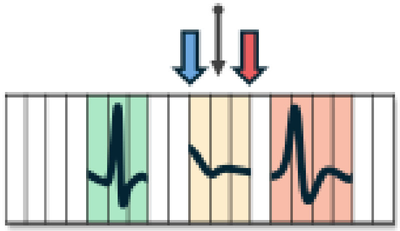
$> \tau$ ✅



$s_2 = (t_2^s, t_2^e)$

$r_\theta(\mathbf{x}_{s_2}, \mathbf{x}_{\mathbf{s}_1})$

$> \tau$ ✅



$s_3 = (t_3^s, t_3^e)$

$r_\theta(\mathbf{x}_{s_3}, \mathbf{x}_{\mathbf{s}_{1:2}})$

$\leq \tau$ ⛔

$\rightarrow$

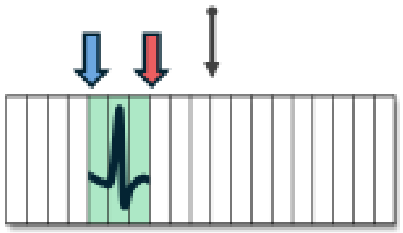
Final Explanation

Method: TimeSeg

6. Stochastic Policy π_ϕ

Factorize Joint Distribution as Step-wise Sampling

$$\pi_\phi(\mathbf{s}_{1:K} \mid \mathbf{x}) = \prod_{k=1}^{K} \pi_\phi(s_k \mid \mathbf{x}, \mathbf{s}_{1:k-1})$$



$$s_k = (t_k^s, t_k^e)$$

Two-step Sampling Process where-to-start, where-to-end

$$s_k = (t_k^s, t_k^e) \sim \pi_\phi(\mathbf{s} \mid \mathbf{x}, \mathbf{s}_{1:k-1}) \stackrel{\text{def}}{=} \pi_{\phi^{\mathrm{s}}}(t^{\mathrm{s}} \mid \mathbf{x}, \mathbf{s}_{1:k-1})\, \pi_{\phi^{\mathrm{e}}}(t^{\mathrm{e}} \mid t^{\mathrm{s}}, \mathbf{x}, \mathbf{s}_{1:k-1}).$$

Final Objectives

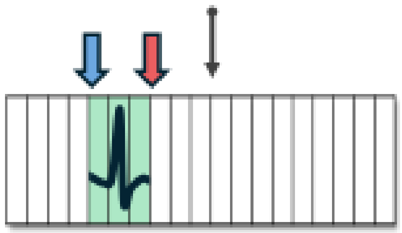
$$\mathcal{L}(\phi) = \mathbb{E}_{\mathbf{x}} \mathbb{E}_{\mathbf{s}_{1:K} \sim \pi_\phi(\cdot \mid \mathbf{x})} \left[\sum_{k=1}^{K} r_\theta\big(\mathbf{x}_{s_k}, \mathbf{x}_{\mathbf{s}_{1:k-1}}\big) - \lambda c\big(s_k\big) \right] \qquad \textbf{(4.7)}$$

Method: TimeSeg

6. Stochastic Policy π_ϕ

Factorize Joint Distribution as Step-wise Sampling

$$\pi_\phi(\mathbf{s}_{1:K} \mid \mathbf{x}) = \prod_{k=1}^{K} \pi_\phi(s_k \mid \mathbf{x}, \mathbf{s}_{1:k-1})$$



$$s_k = (t_k^s, t_k^e)$$

Two-step Sampling Process where-to-start, where-to-end

$$s_k = (t_k^s, t_k^e) \sim \pi_\phi(\mathbf{s} \mid \mathbf{x}, \mathbf{s}_{1:k-1}) \stackrel{\text{def}}{=} \pi_{\phi^{\mathrm{s}}}(t^{\mathrm{s}} \mid \mathbf{x}, \mathbf{s}_{1:k-1})\, \pi_{\phi^{\mathrm{e}}}(t^{\mathrm{e}} \mid t^{\mathrm{s}}, \mathbf{x}, \mathbf{s}_{1:k-1}).$$

Final Objectives

$\mathcal{L}(\phi)$

Why RL?

Joint sampling of (t^s, t^e) with constraint $t^s \leq t^e$ prevents **straightforward reparameterization**

→ **Policy Gradient**

PPO

Experiments

Quantitative · Qualitative · Ablation analysis

Datasets & Benchmarks

5.1 Datasets

Synthetic w/ ground-truth

- SeqComb-UV
- FreqShapes-V
- LowVarDetect-UV

Annotated w/ ground-truth

- MIT-ECG (QRS interval)

Real World w/o ground-truth

- Epilepsy
- Wafer
- GunPoint

5.2 Benchmark Methods

General XAI

- Integrated Gradients*

Point-wise XAI

- Dynamask
- WinIT
- TimeX++*

Patch-wise XAI

- LIMESegment

Evaluation

5.3 Metrics

Overlap w/ GT

- F1-score (↑)

$$\mathrm{F1} = \frac{2 \cdot |\mathbf{m}_{1:K} \cap \mathbf{m}^*|}{|\mathbf{m}_{1:K}| + |\mathbf{m}^*|}$$

- IoU (↑)

$$\mathrm{IoU} = \frac{|\mathbf{m}_{1:K} \cap \mathbf{m}^*|}{|\mathbf{m}_{1:K} \cup \mathbf{m}^*|}$$

Explanation Fidelity w/o GT

- Sufficiency (↓)

$$\mathrm{Suff}\downarrow = 1 - \frac{\mathrm{AUROC}\big(g_\theta(\mathbf{x}_{\mathbf{s}_{1:K}}), y\big)}{\mathrm{AUROC}\big(g_\theta(\mathbf{x}), y\big)}$$

- Comprehensiveness (↑)

$$\mathrm{Comp}\uparrow = 1 - \frac{\mathrm{AUROC}\big(g_\theta(\mathbf{x}^c_{\mathbf{s}_{1:K}}), y\big)}{\mathrm{AUROC}\big(g_\theta(\mathbf{x}), y\big)}$$

Segment Quality

- Contiguity (↓)

$$\mathrm{Cont.} = \frac{1}{T-1} \sum_{t=1}^{T-1} \mathbb{I}\left[m_{1:K,t} \neq m_{1:K,t+1}\right]$$

- Sparsity

$$\mathrm{Sparsity} = \frac{1}{T} \|\mathbf{m}_{1:K}\|_1$$

Reults

5.4 Quantitative Analysis (w/ Ground-Truth)

Table 5.1 Overlap with Ground Truth

Method	SeqComb-UV			LowVarDetect-UV			FreqShapes-V			MIT-ECG		
	F1 ↑	IoU ↑	Cont. ↓	F1 ↑	IoU ↑	Cont. ↓	F1 ↑	IoU ↑	Cont. ↓	F1 ↑	IoU ↑	Cont. ↓
IG*	0.539	0.379	0.166	**0.582**	**0.421**	0.170	0.687	0.534	0.101	0.589	0.435	0.056
Dynamask	0.157	0.088	0.215	0.177	0.115	0.181	0.103	0.060	0.144	0.353	0.221	0.072
WinIT	0.208	0.123	0.178	0.383	0.242	0.117	0.454	0.307	0.114	0.241	0.147	0.133
LIMESegment	0.430	0.289	**0.009**	0.467	0.314	**0.012**	0.440	0.287	**0.030**	0.491	0.359	**0.006**
TimeX++*	0.636	0.489	0.098	0.432	0.291	0.231	**0.799**	**0.666**	0.120	0.593	0.460	0.016
TimeSeg	**0.645**	**0.495**	0.017	0.499	0.356	0.020	0.722	0.576	0.085	**0.739**	**0.621**	0.006

Results

5.4 Quantitative Analysis (w/o Ground-Truth)

Table 5.2 Occlusion Analysis (Mean Substitution)

Method	MIT-ECG			Epilepsy			Wafer			GunPoint		
	Suff. ↓	Comp. ↑	Cont. ↓	Suff. ↓	Comp. ↑	Cont. ↓	Suff. ↓	Comp. ↑	Cont. ↓	Suff. ↓	Comp. ↑	Cont. ↓
Random	55.21	0.01	0.07	3.12	0.04	0.22	34.23	0.01	0.18	32.42	1.42	0.31
WinIT	**23.51**	0.49	0.09	**1.99**	0.09	0.20	49.12	**1.95**	0.05	**32.39**	1.88	0.14
Dynamask	45.31	2.51	0.02	3.06	**0.60**	0.14	**31.13**	1.59	0.05	35.02	**9.15**	0.10
TimeX++*	40.31	**5.31**	**0.01**	6.79	0.07	**0.07**	52.39	0.22	**0.03**	45.41	7.10	**0.01**
TimeSeg	**0.70**	**5.54**	**0.01**	**1.94**	**0.93**	**0.01**	**0.19**	**2.42**	**0.01**	**1.44**	**46.41**	**0.02**

Results

5.4 Quantitative Analysis (w/o Ground-Truth)

Table 5.2 Occlusion Analysis (Mean Substitution)

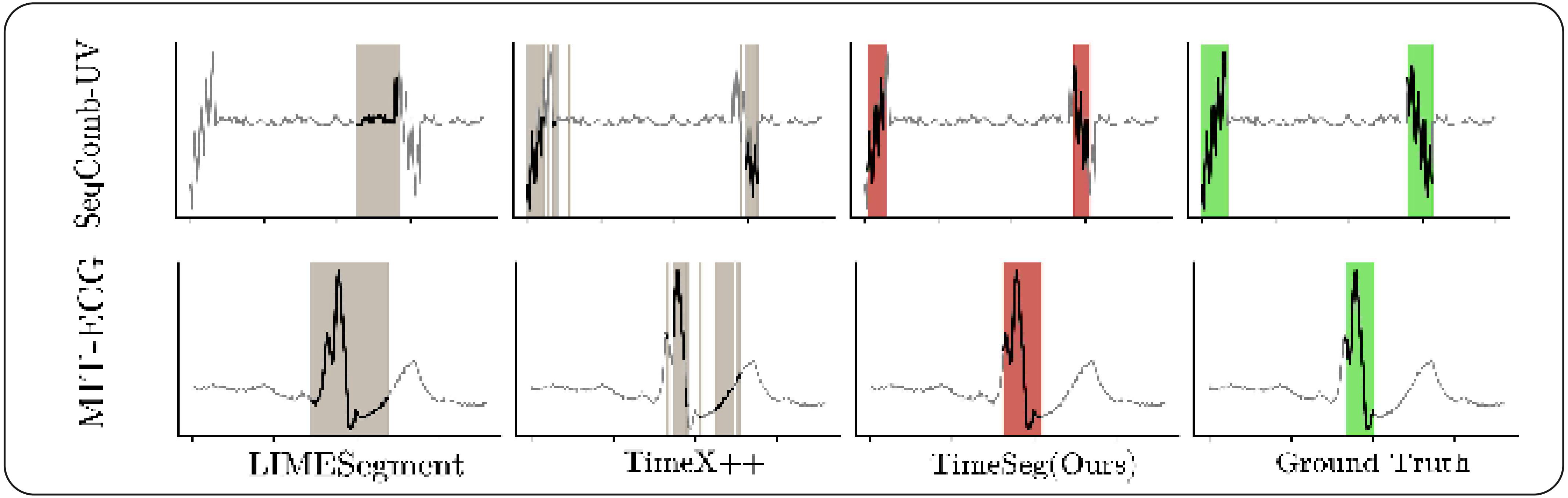
Method	MIT-ECG			Epilepsy			Wafer			GunPoint		
	Suff. ↓	Comp. ↑	Cont. ↓	Suff. ↓	Comp. ↑	Cont. ↓	Suff. ↓	Comp. ↑	Cont. ↓	Suff. ↓	Comp. ↑	Cont. ↓
Random	55.21	0.01	0.07	3.12	0.04	0.22	34.23	0.01	0.18	32.42	1.42	0.31
WinIT	**23.51**	0.49	0.09	**1.99**	0.09	0.20	49.12	**1.95**	0.05	**32.39**	1.88	0.14
Dynamask	45.31	2.51	0.02	3.06	**0.60**	0.14	**31.13**	1.59	0.05	35.02	**9.15**	0.10
TimeX++*	40.31	**5.31**	**0.01**	6.79	0.07	**0.07**	52.39	0.22	**0.03**	45.41	7.10	**0.01**
TimeSeg	**0.70**	**5.54**	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

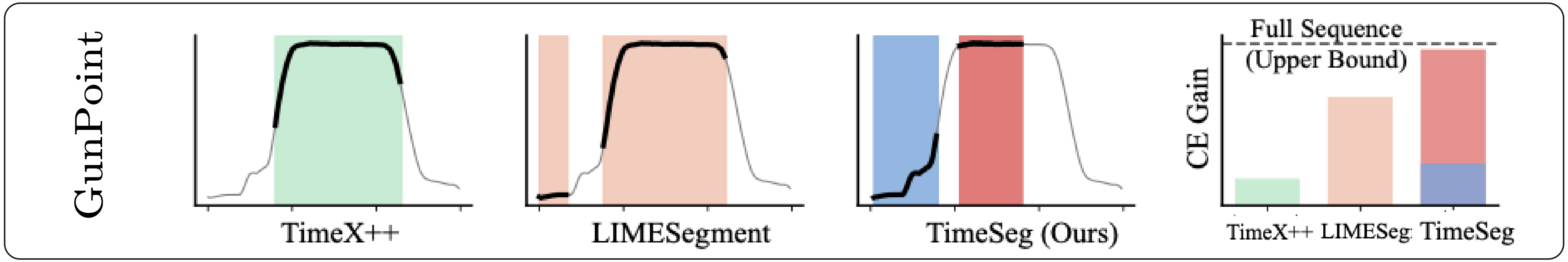
- **TimeSeg** achieves **best** or **second-best** under strict black-box.
- Notably, **Contiguity ≤ 0.02** — concise & well-connected explanations.

Results

5.5 Qualitative Analysis

- Qualitative comparison on **SeqComb-UV, MIT-ECG, and GunPoint**

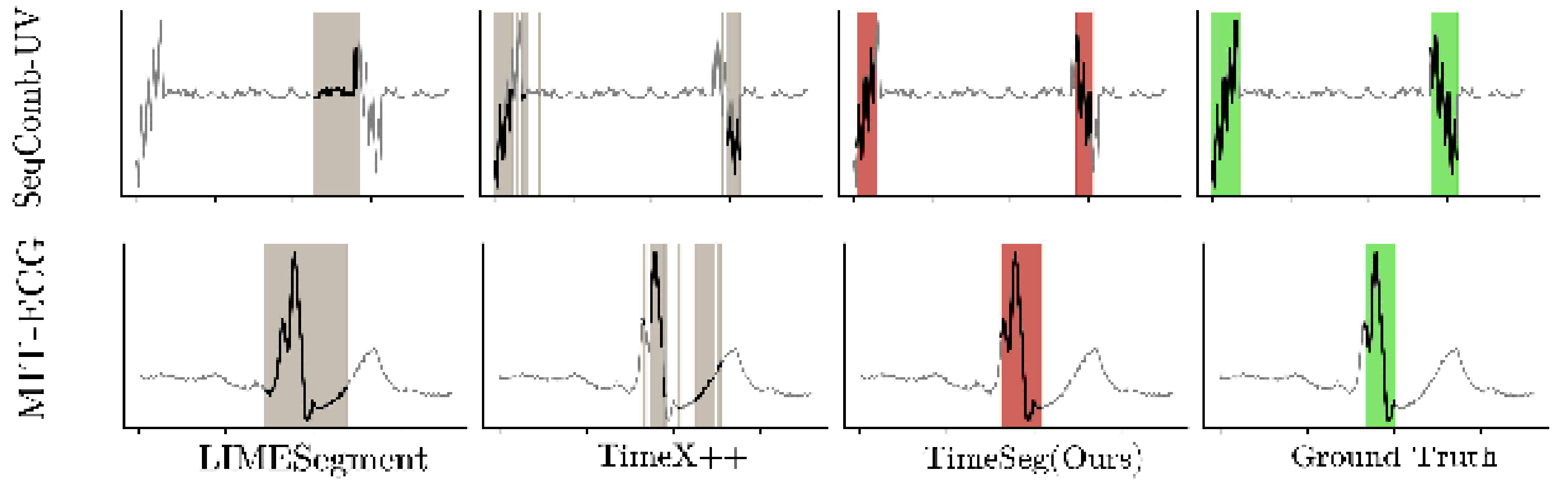




Results

5.5 Qualitative Analysis

- Qualitative comparison on **SeqComb-UV, MIT-ECG, and GunPoint**

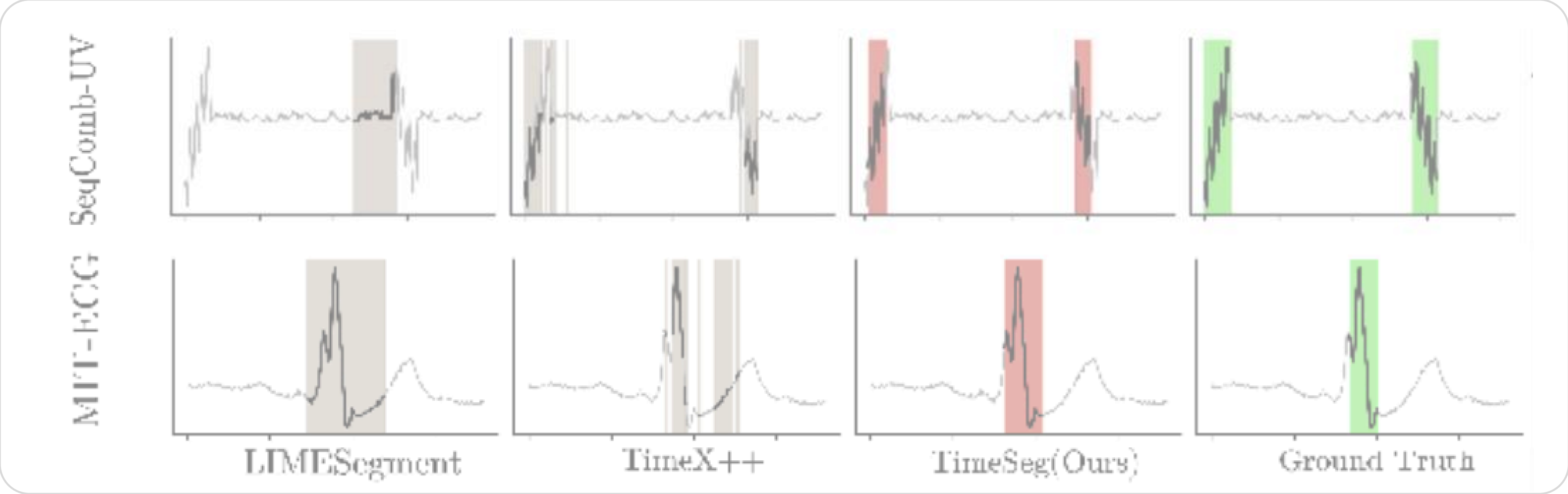


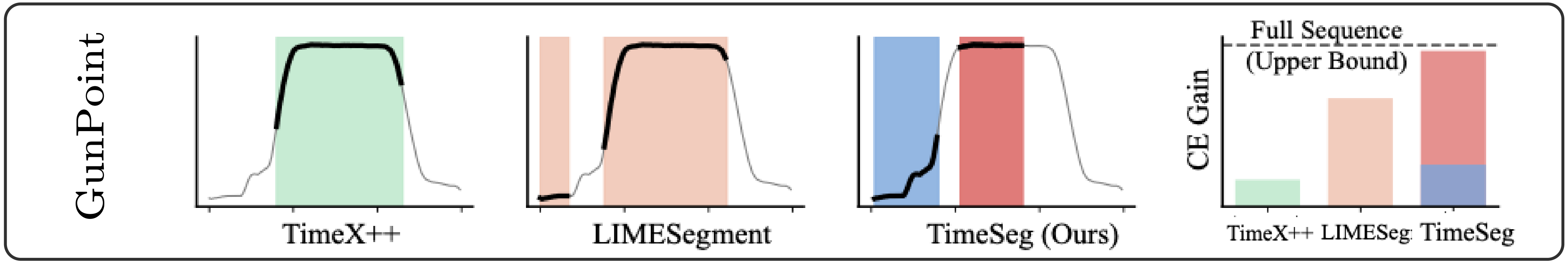
- **TimeX++** — **pinpoints** relevant regions but produces **fragmented**
- **LIMESegment** — **contiguous** but **misaligned** due to fixed-size patches
- **TimeSeg** — **complete, well-aligned segments** matching GT

Results

5.5 Qualitative Analysis

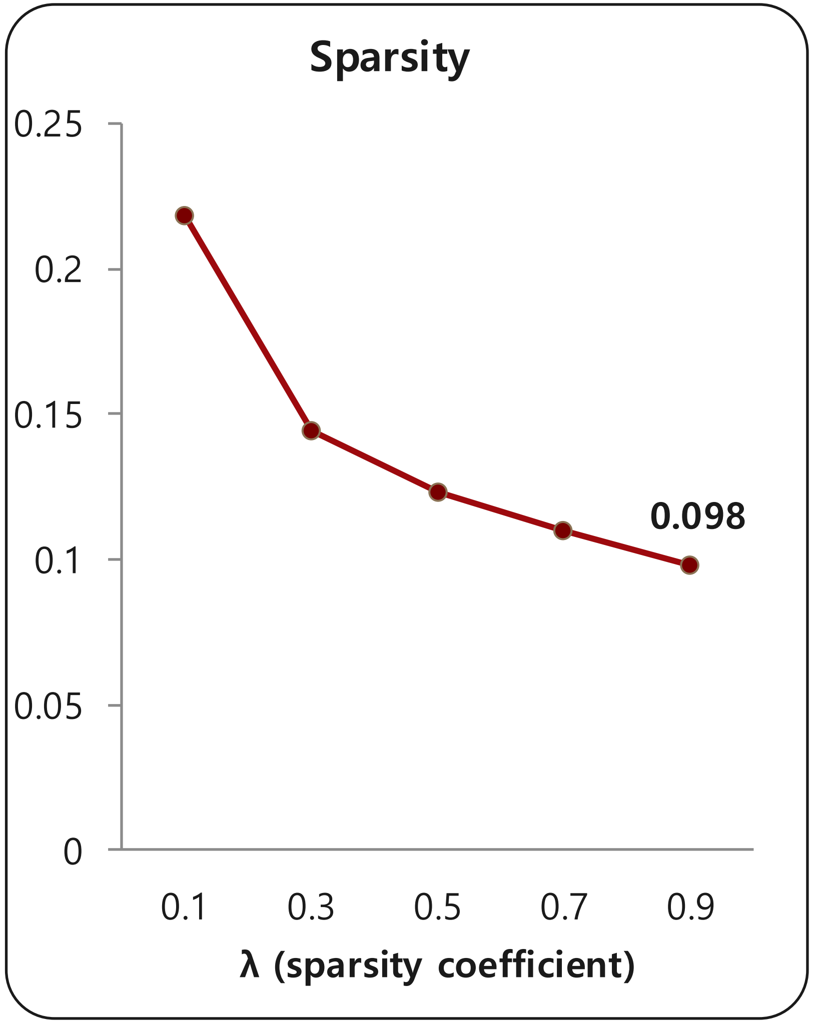
- Qualitative comparison on **SeqComb-UV, MIT-ECG, and GunPoint**

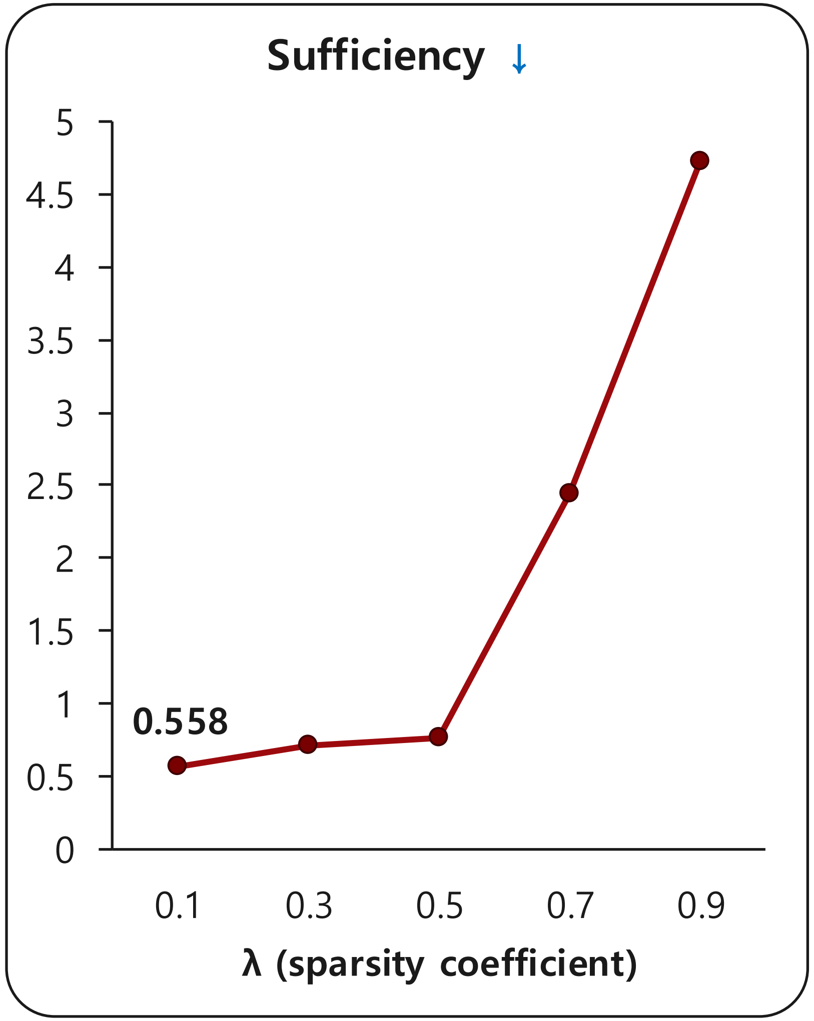


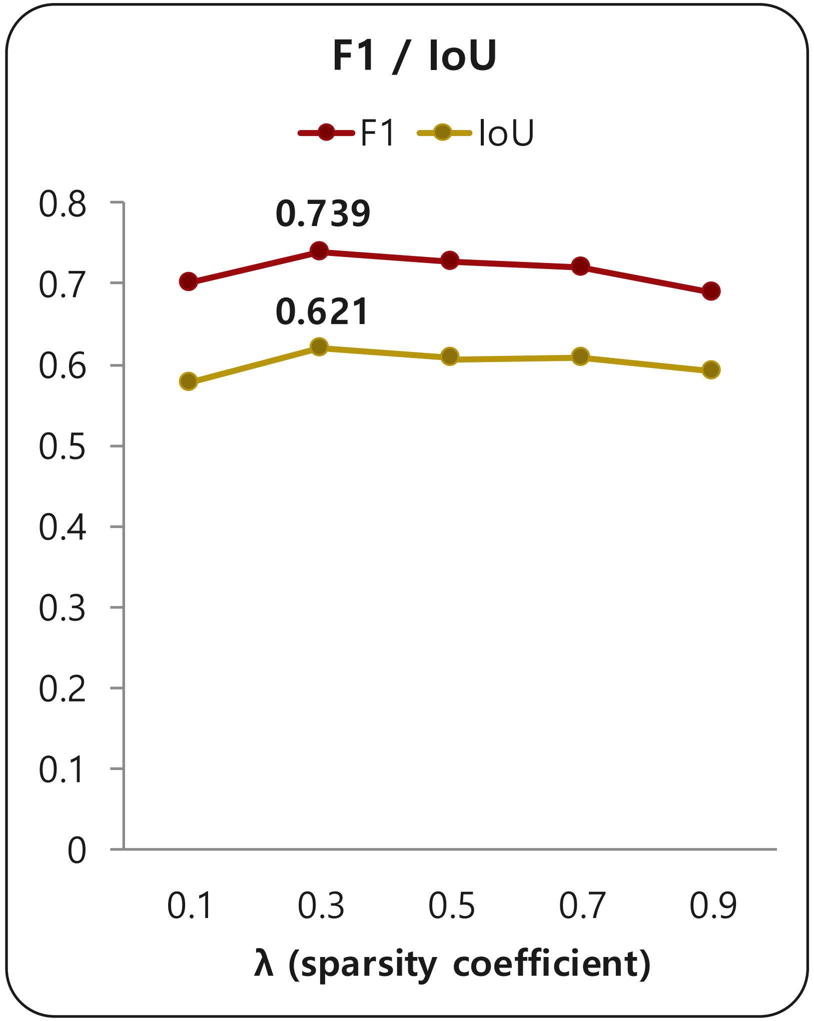


Results

5.6 Ablation Study on λ (MIT-ECG)







Conclusion

Segment-wise Explainer
Problem Formulation

Reformulation to
Sequential Process

Strick
black-box Model

Future Works

- Extending the framework to multivariate time-series by modeling segment selection across chan nels and their inter-channel dependencies

- Extending reward design with human preferences and domain priors for clinically meaningful ex planations

Thank you

Thank you for your attention