Appendices

A Proofs

A.1 Proof of example: $\mathcal{L}_{\text{CE}}$

The minimization problem presented in the example of Section 3.1 can be simplified using, among others, the definition of $\mathcal{L}_{\text{CE}}$ and $\mathbf{y}_i$:

$$\begin{aligned}
&\arg\min_\phi \textstyle\sum_i \mathcal{L}_{\text{CE}}(\phi * \mathbf{x_i}, \mathbf{y}_i) \\
&= \arg\min_\phi \textstyle\sum_i \frac{1}{N} \mathcal{L}_{\text{CE}}(\phi * \mathbf{x_i}, \mathbf{y}_i) \\
&\overset{N \text{ large}}{\approx} \arg\min_\phi \mathbb{E}_{\epsilon\sim E}[\mathcal{L}_{\text{CE}}(\phi * \mathbf{x_i}, \mathbf{y}_i)] \\
&= -\arg\min_\phi \mathbb{E}_{\epsilon\sim E}[\textstyle\sum_t y_t \log((\phi * \mathbf{x}_i)_t) \\
&\qquad\qquad + (1-y_t)\log(1-(\phi * \mathbf{x}_i)_t)] \\
&= \arg\max_\phi \mathbb{E}_{\epsilon\sim E}[\textstyle\sum_t y_t \log((\phi * \mathbf{x}_i)_t) \\
&\qquad\qquad + (1-y_t)\log(1-(\phi * \mathbf{x}_i)_t)] \\
&= \arg\max_\phi \mathbb{E}_{\epsilon_i\sim E}[\textstyle\sum_t \mathbb{1}_{[t=t_i+\epsilon_i]} \log((\phi * \mathbf{x}_i)_t) \\
&\qquad\qquad + (1-\mathbb{1}_{[t=t_i+\epsilon_i]})\log(1-(\phi * \mathbf{x}_i)_t)] \\
&= \arg\max_\phi \mathbb{E}_{\epsilon\sim E}[\log((\phi * \mathbf{x}_i)_{t_i+\epsilon_i}) \\
&\qquad\qquad + \textstyle\sum_{t\neq t_i+\epsilon_i} \log(1-(\phi * \mathbf{x}_i)_t] \\
&= \arg\max_\phi \textstyle\sum_k P(E=k)[\log((\phi * \mathbf{x}_i)_{t_i+k}) \\
&\qquad\qquad + \textstyle\sum_{t\neq t_i+k} \log(1-(\phi * \mathbf{x}_i)_t] \\
&= \arg\max_\phi \textstyle\sum_k P(E+t_i=k)\log((\phi * \mathbf{x}_i)_k) \\
&\qquad\qquad + (1-P(E+t_i=k))\log(1-(\phi * \mathbf{x}_i)_k)
\end{aligned} \tag{A.1}$$

Using the definition of $\mathbf{x}$ and of the convolution operation, the expression can be simplified further:

$$\begin{aligned}
&\arg\max_\phi \textstyle\sum_k P(E+t_i=k)\log((\phi * \mathbf{x}_i)_k) \\
&\qquad + (1-P(E+t_i=k))\log(1-(\phi * \mathbf{x}_i)_k) \\
&= \arg\max_\phi \textstyle\sum_k P(E+t_i=k)\log(\sum_t \phi(t)\cdot x_{i,k-t}) \\
&\qquad + (1-P(E+t_i=k))\log(1-\textstyle\sum_t \phi(t)\cdot x_{i,k-t} \\
&= \arg\max_\phi \textstyle\sum_k P(E+t_i=k)\log(\phi(k-t_i)) \\
&\qquad + (1-P(E+t_i=k))\log(1-\phi(k-t_i)) \\
&= \arg\max_\phi \textstyle\sum_k P(E=k-t_i)\log(\phi(k-t_i)) \\
&\qquad + (1-P(E=k-t_i))\log(1-\phi(k-t_i)) \\
&= \arg\max_\phi \textstyle\sum_k P(E=k)\log(\phi(k)) \\
&\qquad + (1-P(E=k))\log(1-\phi(k)).
\end{aligned} \tag{A.2}$$

Since the value of the different timesteps (k) are mutually independent from one another in the optimization problem, each bin of the convolution filter ϕ can be optimized separately, i.e.,

$$\begin{aligned}
&\arg\max_{\phi(k)} P(E=k)\log(\phi(k)) \\
&\qquad + (1-P(E=k))\log(1-\phi(k)).
\end{aligned} \tag{A.3}$$

Each individual optimization problem can be expressed as

$$\arg\max_x \alpha\log(x) + (1-\alpha)\log(1-x), \tag{A.4}$$

which has the following closed-form solution:

$$\begin{aligned}
&\frac{\partial}{\partial x}\alpha\log(x) + (1-\alpha)\log(1-x) \overset{!}{=} 0 \\
&\quad\iff \frac{\alpha}{x} - \frac{1-\alpha}{1-x} \overset{!}{=} 0 \\
&\quad\overset{x\neq 0,1}{\iff} \alpha(1-x) \overset{!}{=} (1-\alpha)x \\
&\quad\iff x \overset{!}{=} \alpha.
\end{aligned} \tag{A.5}$$

Thus, combining Eq. (A.3) and Eq. (A.5), we obtain the result reported in Eq. (5) from the main text:

$$\begin{aligned}
&\phi^* = \arg\min_\phi \textstyle\sum_i \mathcal{L}_{\text{CE}}(\phi * \mathbf{x_i}, \mathbf{y}_i) \\
&\quad\iff \phi^*(\tau) \approx P(E=\tau), \forall\tau.
\end{aligned} \tag{A.6}$$

A.2 Proof of example: $\mathcal{L}_{\text{LS|CE}}$

The derivation of the result from Eq. (9) is similar to the one presented in Section A.1; thus, a summarized version of the proof is given instead:

$$\begin{aligned}
&\arg\min_\phi \textstyle\sum_i \mathcal{L}_{\text{LS|CE}}(\phi * \mathbf{x_i}, \mathbf{y}_i \,|\, \Phi) \\
&= \arg\min_\phi \textstyle\frac{1}{N}\sum_i \mathcal{L}_{\text{LS|CE}}(\phi * \mathbf{x_i}, \mathbf{y}_i \,|\, \Phi) \\
&\overset{N \text{ large}}{\approx} \arg\min_\phi \mathbb{E}_{\epsilon\sim E}[\mathcal{L}_{\text{LS|CE}}(\phi * \mathbf{x_i}, \mathbf{y}_i \,|\, \Phi)] \\
&= \arg\min_\phi \mathbb{E}_{\epsilon\sim E}[-\textstyle\sum_t \Big((\Phi * \mathbf{y}_i)_t \log((\phi * \mathbf{x}_i)_t) \\
&\qquad\qquad + (1-(\Phi * \mathbf{y}_i)_t \log(1-(\phi * \mathbf{x}_i)_t) \Big)] \\
&= \arg\max_\phi \mathbb{E}_{\epsilon\sim E}[\textstyle\sum_t \Big((\Phi * \mathbf{y}_i)_t \log((\phi * \mathbf{x}_i)_t) \\
&\qquad\qquad + (1-(\Phi * \mathbf{y}_i)_t \log(1-(\phi * \mathbf{x}_i)_t) \Big)] \\
&= \arg\max_\phi \mathbb{E}_{\epsilon\sim E}[\textstyle\sum_t \Big(\sum_{\tau=0}^T y_{i,\tau}\Phi(t-\tau)\log((\phi * \mathbf{x}_i)_t) \\
&\qquad\qquad + (1-\textstyle\sum_{\tau=0}^T y_{i,\tau}\Phi(t-\tau))\log(1-(\phi * \mathbf{x}_i)_t) \Big)] \\
&= \arg\max_\phi \mathbb{E}_{\epsilon\sim E}[\textstyle\sum_t \Big(\Phi(t-t_i-\epsilon_i)\log((\phi * \mathbf{x}_i)_t) \\
&\qquad\qquad + (1-\Phi(t-t_i-\epsilon_i))\log(1-(\phi * \mathbf{x}_i)_t) \Big)] \\
&= \arg\max_\phi \textstyle\sum_k P(E=k)\sum_t \Big(\Phi(t-t_i-\epsilon_i)\log((\phi * \mathbf{x}_i)_t) \\
&\qquad\qquad + (1-\Phi(t-t_i-\epsilon_i))\log(1-(\phi * \mathbf{x}_i)_t) \Big) \\
&= \arg\max_\phi \textstyle\sum_t\sum_k P(E=k) \Big(\Phi(t-t_i-\epsilon_i)\log((\phi * \mathbf{x}_i)_t) \\
&\qquad\qquad + (1-\Phi(t-t_i-\epsilon_i))\log(1-(\phi * \mathbf{x}_i)_t) \Big)] \\
&= \arg\max_\phi \textstyle\sum_t (E * \Phi)_{t-t_i}\log((\phi * \mathbf{x}_i)_t) \\
&\qquad\qquad + (E * (1-\Phi))_{t-t_i}\log(1-(\phi * \mathbf{x}_i)_t) \\
&= \arg\max_\phi \textstyle\sum_t (E * \Phi)_{t-t_i}\log((\phi * \mathbf{x}_i)_t) \\
&\qquad\qquad + (1-(E * \Phi)_{t-t_i})\log(1-\hat{p}_{\theta,t})
\end{aligned} \tag{A.7}$$

Thus, using Eq. (A.5) and the same argument as in Section A.1, we obtain the final result (Eq. (8) in the main text):

$$\begin{aligned}
&\phi^* = \arg\min_\phi \textstyle\sum_i \mathcal{L}_{\text{LS|CE}}(\phi * \mathbf{x_i}, \mathbf{y}_i \,|\, \Phi) \\
&\quad\iff \phi^*(\tau) \approx (E * \Phi)_\tau = \textstyle\sum_i P(E=i)\Phi(\tau - i), \forall\tau.
\end{aligned} \tag{A.8}$$

A.3 Proof of example: $\mathcal{L}_{\text{SLL}}$

The beginning of the derivation is done as in previous sections:

$$
\begin{aligned}
&\arg\min_\phi \textstyle\sum_i \mathcal{L}_{\text{SLL}}(\phi * \mathbf{x_i}, \mathbf{y}_i \,|\, \Phi, \mathcal{E}) \\
&= \arg\min_\phi \tfrac{1}{N} \textstyle\sum_i \mathcal{L}_{\text{SLL}}(\phi * \mathbf{x_i}, \mathbf{y}_i \,|\, \Phi, \mathcal{E}) \\
&\overset{N \text{ large}}{\approx} \arg\min_\phi \mathbb{E}_{\epsilon\sim E}[\mathcal{L}_{\text{SLL}}(\phi * \mathbf{x_i}, \mathbf{y}_i \,|\, \Phi, \mathcal{E})] \\
&= \arg\min_\phi \mathbb{E}_{\epsilon\sim E}[\textstyle\sum_t \Big((\mathcal{E} * \Phi * \phi * \mathbf{x}_i)_t - (\Phi * \mathbf{y}_i)_t \Big)^2] \\
&= \arg\min_\phi \mathbb{E}_{\epsilon\sim E}[\textstyle\sum_t (\mathcal{E} * \Phi * \phi * \mathbf{x}_i)_t^2 \\
&\qquad\qquad +(\Phi * \mathbf{y}_i)_t^2 \\
&\qquad\qquad -2(\mathcal{E} * \Phi * \phi * \mathbf{x}_i)_t \cdot (\Phi * \mathbf{y}_i)_t] \\
&= \arg\min_\phi \mathbb{E}_{\epsilon_i\sim E}[\textstyle\sum_t (\mathcal{E} * \Phi * \phi * \mathbf{x}_i)_t^2 \\
&\qquad\qquad +\Phi(t-t_i-\epsilon_i)^2 \\
&\qquad\qquad -2(\mathcal{E} * \Phi * \phi * \mathbf{x}_i)_t \cdot \Phi(t-t_i-\epsilon_i)] \\
&= \arg\min_\phi \textstyle\sum_t (\mathcal{E} * \Phi * \phi * \mathbf{x}_i)_t^2 \\
&\qquad +\mathbb{E}_{\epsilon_i\sim E}[\textstyle\sum_t \Phi(t-t_i-\epsilon_i)^2 \\
&\qquad\qquad -2(\mathcal{E} * \Phi * \phi * \mathbf{x}_i)_t \cdot \Phi(t-t_i-\epsilon_i)] \\
&= \arg\min_\phi \textstyle\sum_t (\mathcal{E} * \Phi * \phi * \mathbf{x}_i)_t^2 \\
&\qquad +\textstyle\sum_k P(E{=}k)[\sum_t \Phi(t-t_i-k)^2 \\
&\qquad\qquad -2(\mathcal{E} * \Phi * \phi * \mathbf{x}_i)_t \cdot \Phi(t-t_i-k)] \\
&= \arg\min_\phi \textstyle\sum_t (\mathcal{E} * \Phi * \phi * \mathbf{x}_i)_t^2 \\
&\qquad +\textstyle\sum_{k,t} P(E{=}k)\Phi(t-t_i-k)^2 \\
&\qquad -2\textstyle\sum_{k,t} P(E{=}k)(\mathcal{E} * \Phi * \phi * \mathbf{x}_i)_t \cdot \Phi(t-t_i-k) \\
&= \arg\min_\phi \textstyle\sum_t (\mathcal{E} * \Phi * \phi * \mathbf{x}_i)_t^2 \\
&\qquad\quad +\textstyle\sum_t (\Phi^2 * E)_{t-t_i} \\
&\qquad\quad -\textstyle\sum_t 2(\mathcal{E} * \Phi * \phi * \mathbf{x}_i)_t \cdot (\Phi * E)_{t-t_i}
\end{aligned}
\tag{A.9}
$$

The second term does not depend on ϕ, thus

$$
\begin{aligned}
&\arg\min_\phi \textstyle\sum_i \mathcal{L}_{\text{SLL}}(\phi * \mathbf{x_i}, \mathbf{y}_i \,|\, \Phi, \mathcal{E}) \\
&\approx \arg\min_\phi \textstyle\sum_t (\mathcal{E} * \Phi * \phi * \mathbf{x}_i)_t^2 \\
&\qquad\qquad -2(\mathcal{E} * \Phi * \phi * \mathbf{x}_i)_t \cdot (\Phi * E)_{t-t_i}
\end{aligned}
\tag{A.10}
$$

Differentiating by ϕ yields,

$$
\begin{aligned}
&\frac{\partial}{\partial \phi} \textstyle\sum_t (\mathcal{E} * \Phi * \phi * \mathbf{x}_i)_t^2 - 2(\mathcal{E} * \Phi * \phi * \mathbf{x}_i)_t \cdot (\Phi * E)_{t-t_i} \\
&\quad = \textstyle\sum_t 2(\mathcal{E} * \Phi * \phi * \mathbf{x}_i)_t \cdot (\mathcal{E} * \Phi * 1 * \mathbf{x}_i)_t \\
&\qquad -2(\mathcal{E} * \Phi * 1 * \mathbf{x}_i)_t \cdot (\Phi * E)_{t-t_i} \\
&\quad \overset{1*\Phi=1}{=} \textstyle\sum_t 2(\mathcal{E} * \Phi * \phi * \mathbf{x}_i)_t \cdot (\mathcal{E} * 1 * \mathbf{x}_i)_t \\
&\qquad -2(\mathcal{E} * 1 * \mathbf{x}_i)_t \cdot (\Phi * E)_{t-t_i} \\
&\quad \overset{1*\mathcal{E}=1}{=} \textstyle\sum_t 2(\mathcal{E} * \Phi * \phi * \mathbf{x}_i)_t \cdot (1 * \mathbf{x}_i)_t \\
&\qquad -2(1 * \mathbf{x}_i)_t \cdot (\Phi * E)_{t-t_i} \\
&\quad \overset{1*\mathcal{E}=1}{=} \textstyle\sum_t 2(\mathcal{E} * \Phi * \phi * \mathbf{x}_i)_t - 2(\Phi * E)_{t-t_i}
\end{aligned}
\tag{A.11}
$$

Using the definition of $\mathbf{x}_i$

$$
\begin{aligned}
&\textstyle\sum_t 2(\mathcal{E} * \Phi * \phi * \mathbf{x}_i)_t - 2(\Phi * E)_{t-t_i} = 0 \\
&\quad \iff \textstyle\sum_t (\mathcal{E} * \Phi * \phi * \mathbf{x}_i)_t - (\Phi * E)_{t-t_i} = 0 \\
&\quad \iff \textstyle\sum_t (\mathcal{E} * \Phi * \mathbf{p}^*)_t - (\Phi * E)_{t-t_i} = 0 \\
&\quad \iff \textstyle\sum_t (\mathcal{E} * \mathbf{p}^*)_t - (E)_{t-t_i} = 0 \\
&\quad \iff \textstyle\sum_t (\mathcal{E} * \mathbf{p}^*)_t - (E{*}\mathbf{g}_i)_t = 0
\end{aligned}
\tag{A.12}
$$

Thus, we obtain the final result presented in Eq. (12), which states that the optimal prediction $\hat{\mathbf{p}}_i^*$ that minimizes the loss $\sum_i \mathcal{L}_{\mathbf{SLL}}(\phi * \mathbf{x_i}, \mathbf{y}_i \,|\, \Phi, \mathcal{E})$ has the form:

$$
(\mathcal{E} {*} \mathbf{p}_i^*)_\tau \approx (E {*} \mathbf{g}_i)_\tau \tag{A.13}
$$

A.4 Proof of example: $\mathcal{L}_{\mathcal{S}\textbf{oftLoc}}$

As done previously,

$$
\begin{aligned}
&\arg\min_\phi \textstyle\sum_i \mathcal{L}_{\mathcal{S}\text{oftLoc}}(\phi * \mathbf{x_i}, \mathbf{y}_i) \\
&\quad = \arg\min_\phi \tfrac{1}{N} \textstyle\sum_i \mathcal{L}_{\mathcal{S}\text{oftLoc}}(\phi * \mathbf{x_i}, \mathbf{y}_i) \\
&\quad \overset{N \text{ large}}{\approx} \arg\min_\phi \mathbb{E}_{\epsilon\sim E}[\mathcal{L}_{\mathcal{S}\text{oftLoc}}(\phi * \mathbf{x_i}, \mathbf{y}_i)]
\end{aligned}
\tag{A.14}
$$

As the training progresses (i.e., $\alpha_\tau \to 1$), the counting-based constraint becomes more predominant and ensures that only one timestep is assigned all the mass for the event of interest. Thus, given the definition of $\mathbf{x}_i$ and ϕ, the smoothed prediction is of the form

$$
(\Phi * \phi * \mathbf{x}_i)_t = \Phi(t - t_i - \beta), \tag{A.15}
$$

where β is a model constant in $\mathbb{N}$. Therefore, as the counting-based constraint already ensure that exactly one timestep has probability 1, while all other are assigned zero probability, it simply remains to show that the model bias β is equal to zero. Indeed, it would apply that the predictions are perfectly aligned with the ground-truth without any bias.

Plugging Eq. (A.14) into Eq. (A.15),

$$
\begin{aligned}
&\arg\min_\phi \textstyle\sum_i \mathcal{L}_{\mathcal{S}\text{oftLoc}}(\phi * \mathbf{x}^{(i)}, \mathbf{y}^{(i)}) \\
&\overset{\alpha_\tau \to 1}{\approx} \arg\min_\beta \mathbb{E}_{\epsilon\sim E}[\textstyle\sum_t \Big(\Phi(t-t^{(i)}-\beta) - \Phi(t-t^{(i)}-\epsilon) \Big)^2] \\
&= \arg\min_\beta \mathbb{E}_{\epsilon\sim E}[\textstyle\sum_t \Big(\Phi(t-\beta) - \Phi(t-\epsilon) \Big)^2]
\end{aligned}
\tag{A.16}
$$

Using the properties of the filter Φ, it can be shown that the sum inside the expectation of Equation (A.16) is only a function of the distance between the prediction and the misaligned label (i.e., $|\epsilon - \beta|$):

$$
\begin{aligned}
&\textstyle\sum_t \Big(\Phi(t-\beta) - \Phi(t-\epsilon) \Big)^2 = \begin{cases} \textstyle\sum_{\bar{t}} \Big(\Phi(\bar{t}) - \Phi(t-(\epsilon-\beta)) \Big)^2 \\ \textstyle\sum_{\bar{t}} \Big(\Phi(\bar{t}-(\beta-\epsilon)) - \Phi(t) \Big)^2 \end{cases} \\
&\implies \textstyle\sum_t \Big(\Phi(t-\beta) - \Phi(t-\epsilon) \Big)^2 = \textstyle\sum_t \Big(\Phi(t) - \Phi(t-|\epsilon-\beta|) \Big)^2.
\end{aligned}
\tag{A.17}
$$

Thus, by setting

$$
\gamma(x) := \textstyle\sum_t \Big(\Phi(t) - \Phi(t-|x|) \Big)^2, \tag{A.18}
$$

Equation (A.16) becomes

$$
\begin{aligned}
&\arg\min_\phi \textstyle\sum_i \mathcal{L}_{\mathcal{S}\text{oftLoc}}(\phi * \mathbf{x}^{(i)}, \mathbf{y}^{(i)}) \\
&= \arg\min_\beta \mathbb{E}_{\epsilon\sim E}[\textstyle\sum_t \Big(\Phi(t-\beta) - \Phi(t-\epsilon) \Big)^2] \\
&= \arg\min_\beta \mathbb{E}_{\epsilon\sim E}[\gamma(\epsilon - \beta)] \\
&= \arg\min_\beta \textstyle\sum_k P(E{=}k) \cdot \gamma(k - \beta),
\end{aligned}
\tag{A.19}
$$

where $\gamma(x)$ is a positive and symmetric (around 0) function, which is monotonically increasing in $[0,\infty]$—and ~~thus~~ monotonically decreasing in $[-\infty,0]$ by symmetry.

The main results can thus be proven by showing that

$$\begin{aligned}&\textstyle\sum_k P(E=k)\cdot\gamma(k-\beta)\\ &\qquad\geq \textstyle\sum_k P(E=k)\cdot\gamma(k-0), \forall\beta\in\mathbb{N}.\end{aligned} \tag{A.20}$$

Let us assume that β is *odd*, then the sum can be reordered as follows:

$$\begin{aligned}&\textstyle\sum_k P(E=k)\cdot\gamma(k-\beta)\\ &\quad=\textstyle\sum_{t\geq 0}\Big(P(E=\lceil\beta/2\rceil+t)\cdot\gamma(\lceil\beta/2\rceil+t-\beta)\\ &\qquad\qquad+P(E=\lfloor\beta/2\rfloor-t)\cdot\gamma(\lfloor\beta/2\rfloor-t-\beta)\Big).\end{aligned} \tag{A.21}$$

Using the fact that $\gamma(x)$ is monotonically increasing in $[0,\infty]$ and monotonically decreasing in $[-\infty,0]$ and that the noise distribution is well-behaved (i.e., $P(E=k)\leq P(E=\tilde{k}) \iff |k|\geq|\tilde{k}|$) and symmetric, the following inequality holds:

$$\begin{aligned}&\textstyle\sum_k P(E=k)\cdot\gamma(k-\beta)\\ &\quad=\textstyle\sum_{t\geq 0}\Big(P(E=\lceil\beta/2\rceil+t)\cdot\gamma(\lceil\beta/2\rceil+t-\beta)\\ &\qquad\qquad+P(E=\lfloor\beta/2\rfloor-t)\cdot\gamma(\lfloor\beta/2\rfloor-t-\beta)\Big)\\ &\quad\geq\textstyle\sum_{t\geq 0}\Big(P(E=\lceil\beta/2\rceil+t)\cdot\gamma(\lfloor\beta/2\rfloor-t-\beta)\Big)\\ &\qquad\qquad+P(E=\lfloor\beta/2\rfloor-t)\cdot\gamma(\lceil\beta/2\rceil+t-\beta)\end{aligned} \tag{A.22}$$

The equation can be simplified further using the definition of the rounding operator:

$$\begin{aligned}&\textstyle\sum_k P(E=k)\cdot\gamma(k-\beta)\\ &\quad\geq\textstyle\sum_{t\geq 0}\Big(P(E=\lceil\beta/2\rceil+t)\cdot\gamma(\lfloor\beta/2\rfloor-t-\beta)\Big)\\ &\qquad\qquad+P(E=\lfloor\beta/2\rfloor-t)\cdot\gamma(\lceil\beta/2\rceil+t-\beta)\\ &\quad=\textstyle\sum_{t\geq 0}\Big(P(E=\lceil\beta/2\rceil+t)\cdot\gamma(-\lceil\beta/2\rceil-t)\Big)\\ &\qquad\qquad+P(E=\lfloor\beta/2\rfloor-t)\cdot\gamma(-\lfloor\beta/2\rfloor+t).\end{aligned} \tag{A.23}$$

Finally, the final inequality is obtained by both using the symmetry—around zero—of the function $\gamma(x)$ and performing a reordering of the sum:

$$\begin{aligned}&\textstyle\sum_k P(E=k)\cdot\gamma(k-\beta)\\ &\quad\geq\textstyle\sum_{t\geq 0}\Big(P(E=\lceil\beta/2\rceil+t)\cdot\gamma(-\lceil\beta/2\rceil-t)\Big)\\ &\qquad\qquad+P(E=\lfloor\beta/2\rfloor-t)\cdot\gamma(-\lfloor\beta/2\rfloor+t)\\ &\quad=\textstyle\sum_{t\geq 0}\Big(P(E=\lceil\beta/2\rceil+t)\cdot\gamma(\lceil\beta/2\rceil+t)\Big)\\ &\qquad\qquad+P(E=\lfloor\beta/2\rfloor-t)\cdot\gamma(\lfloor\beta/2\rfloor-t)\\ &\quad=\textstyle\sum_k P(E=k)\cdot\gamma(k).\end{aligned} \tag{A.24}$$

The derivation for β *even* is analogous.

In conclusion, since

$$\begin{aligned}&\textstyle\sum_k P(E=k)\cdot\gamma(k-\beta)\\ &\qquad\geq\textstyle\sum_k P(E=k)\cdot\gamma(k-0), \forall\beta\in\mathbb{N},\end{aligned} \tag{A.25}$$

then the main result

$$\begin{aligned}\beta^* &= \arg\min_\beta \textstyle\sum_{t,k} P(E=k)\Big(\Phi(t-\beta)-\Phi(t-k)\Big)^2\\ &\implies \beta^*=0\end{aligned} \tag{A.26}$$

follows from Equation (A.19). Of course, stronger statements—with weaker assumptions—could be derived if the noise distribution E was explicitly known (e.g., $E=N(0,\sigma^2)$).

A.5 Proof of Counting Sparsity Lemma

The statement to prove is the following:

$$D_{KL}(\mathbb{1}_c\|\textstyle\sum_i\mathfrak{B}(\hat{p}_{\theta,i}))=0\Leftrightarrow\ell_0(\hat{\mathbf{p}}_\theta)=c\wedge\hat{\mathbf{p}}_\theta\in\{0,1\}^T \tag{A.27}$$

We prove this equivalence (i.e., $P\iff Q$) by showing that each conditional holds (i.e., $P\Rightarrow Q\wedge Q\Rightarrow P$)

$\Rightarrow$

$$\begin{aligned}&D_{KL}(\mathbb{1}_c\|\textstyle\sum_i\mathfrak{B}(\hat{p}_{\theta,i}))=0\Rightarrow\begin{cases}\mu(\mathbb{1}_c)=\mu(\sum_i\mathfrak{B}(\hat{p}_{\theta,i}))\\ \sigma^2(\mathbb{1}_c)=\sigma^2(\sum_i\mathfrak{B}(\hat{p}_{\theta,i}))\end{cases}\\ &\quad\iff\begin{cases}c=\sum_i\hat{p}_{\theta,i}\\ 0=\sum_i(1-\hat{p}_{\theta,i})\hat{p}_{\theta,i}\end{cases}\\ &\quad\iff\begin{cases}c=\sum_i\hat{p}_{\theta,i}\\ \hat{p}_{\theta,i}\in\{0,1\}\end{cases}\\ &\quad\Rightarrow\ell_0(\hat{\mathbf{p}}_\theta)=c\wedge\hat{\mathbf{p}}_\theta\in\{0,1\}^T\end{aligned} \tag{A.28}$$

$\Leftarrow$

$$\begin{aligned}&\ell_0(\hat{\mathbf{p}}_\theta)=c\wedge\hat{\mathbf{p}}_\theta\in\{0,1\}^T\\ &\quad\Rightarrow\textstyle\sum_i\mathfrak{B}(\hat{p}_{\theta,i}) = \sum_{\{i|\hat{p}_{\theta,i}=1\}}\mathbb{1}_1+\sum_{\{i|\hat{p}_{\theta,i}=0\}}0\\ &\qquad\qquad=\textstyle\sum_{i=1}^c\mathbb{1}_1=\mathbb{1}_c\end{aligned} \tag{A.29}$$

B Additional Experiments

B.1 Golf Video Event Detection Experiment

Full Results Table B.1 presents the full results of the golf swing sequencing experiment described in Section 5.1. This experiment not only reveals that the results are overall consistent across architectures (i.e., unidirectional vs. bidirectional model), but also highlights that our loss function significantly outperforms all other benchmarks in terms of robustness to label misalignment.

This performance gap could be partially explained by the significant difference in the temporal ambiguity of the predictions yielded by the various models. Indeed, Figure B.1—which displays the event occurrence probabilities inferred by the different models for a given test sequence—confirms the theoretical claim made in Section 3.2 that label smoothing-based models lead to temporally dispersed predictions, while our approach infers more clear-cut temporal locations. (Note that even more clear-cut point predictions could be achieved for $\mathcal{L}_{\mathcal{S}\text{oftLoc}}$ by training the model further than the 10k iterations set by (McNally et al. 2019) since it would allow for full convergence of the counting loss function.) The detrimental impact of increased prediction ambiguity on the performance metrics is exacerbated by the small temporal error tolerance set for the task (McNally et al. 2019).

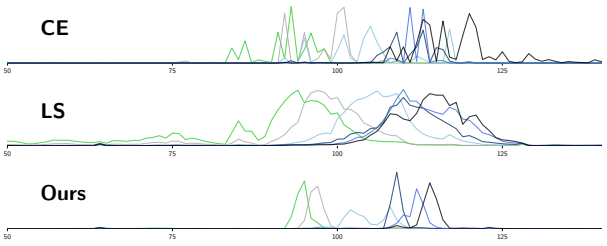


Figure B.1: Out-of-Sample Golf Swing Action Predictions. **Ours**: sharp predictions, **LS** (label smoothing): ambiguous predictions, **CE**: multiple peaks. (Test sequence: 0, split: 1, noise level: $\sigma = 3$ frames.)

Ablation Study Table B.2 reveals that the counting regularization consistently improves the *robustness* (to label misalignment) of the golf swing event sequencing model on this experiment. Indeed, training with this regularizer yields half a percent improvement in accuracy on all noise levels (with the exception of the noise-free case) when compared to the results obtained when training with $\mathcal{L}_{\text{SLL}}$ alone. This increase represents a non-negligible improvement since the particular task introduced by (McNally et al. 2019) does not necessarily require complex post-processing operations to be solved successfully. Indeed, the sequences in the dataset contain exactly one occurrence for each event class (see Section 5.1 for more details about the characteristics of the golf video sequencing dataset). Thus, in this setup, a simple global maximum-picking operation can be used to achieve detection sparsity since the model does not require multi-instance detection capabilities. Overall, this result highlights that the incorporation of prediction sparsity directly in the learning process through count-based regularization can improve the model's detection ability even in settings where prediction sparsity is easily achievable.

B.2 Wearable Sensor Experiment

Other Noise Distributions As mentioned in Section 5.2, the smoking puff detection experiment (Adams and Marlin 2017) has not only been conducted with a normal distribution of label misalignment (see Table 2), but also with more challenging noise patterns: binary constant length shifting of labels ($\pm\delta$ steps with equal probability) denoted $\mathcal{B}(-\delta, \delta)$

Table B.2: Golf Action Detection Ablation Study. Performance comparison of $\mathcal{L}_{\text{SLL}}$ and $\mathcal{L}_{\mathcal{S}\text{oftLoc}}$ using a unidirectional RNN (McNally et al. 2019) with respect to label misalignment distribution $\lfloor \mathcal{N}(0, \sigma^2) \rceil$. ($\sigma$ in frames). The (4-fold) cross-validated **mean accuracy** is reported.

	$\sigma=0$	1	2	3	4
$\mathcal{L}_{\text{SLL}}$	**63.4**	61.6	58.6	53.9	50.1
$\mathcal{L}_{\mathcal{S}\text{OFTLOC}}$	63.0	**62.2**	**59.3**	**54.9**	**50.7**

Table B.1: Golf Action Detection Full Result. Performance comparison of various training losses ($\mathcal{L}_{\text{CE}}$, $\mathcal{L}_{\text{LS|CE}}$, $\mathcal{L}_{\text{LS|SE}}$, and $\mathcal{L}_{\mathcal{S}\text{oftLoc}}$) on the golf swing sequencing task (McNally et al. 2019) with respect to various label misalignment levels $\lfloor \mathcal{N}(0, \sigma^2) \rceil$ (σ in number of frames). The cross-validated (4-folds) **mean accuracy** is reported.

(a) Bidirectional RNN

	$\sigma=0$	1	2	3	4
$\mathcal{L}_{\text{CE}}$	68.1	60.4	51.6	43.1	36.9
$\mathcal{L}_{\text{LS|CE}}$	66.7	64.7	59.1	54.8	49.1
$\mathcal{L}_{\text{LS|SE}}$	**69.1**	66.2	60.6	54.7	50.7
$\mathcal{L}_{\mathcal{S}\text{OFTLOC}}$	67.2	**68.0**	**65.6**	**58.6**	**54.2**

(b) Unidirectional RNN (i.e., *causal model*)

	$\sigma=0$	1	2	3	4
$\mathcal{L}_{\text{CE}}$	62.8	57.2	47.3	40.9	35.3
$\mathcal{L}_{\text{LS|CE}}$	57.0	54.2	50.6	46.4	42.5
$\mathcal{L}_{\text{LS|SE}}$	61.3	59.5	55.2	49.9	46.5
$\mathcal{L}_{\mathcal{S}\text{OFTLOC}}$	**63.0**	**62.2**	**59.3**	**54.9**	**50.7**

and skewed-normal noise distribution $\mathcal{SN}(0, \sigma^2, \alpha = -2)$. The full results are summarized in Table B.3. Overall, our approach displays very consistent improvements over all other tested benchmarks. Indeed, the model trained with $\mathcal{L}_{\mathcal{S}\text{oftLoc}}$ not only outperforms all benchmarks on all noise levels and all misalignment distributions, but also yields the results with the least variability (i.e., the lowest standard deviation of performance). In addition, the substantial performance gap for higher noise levels (e.g., $\sigma, \delta = 4$) further demonstrates the effectiveness of our novel loss function.

Ablation Study As done previously, we perform an ablation study to measure the impact of the counting-based sparsity regularization on the performance of the smoking puff time series detection model. The results are summarized in Table B.4. Overall, adding the Poisson-binomial loss functions as loss as a regularizer to the soft localization learning loss consistently improves the performance of the trained model. This result is in line with the findings of the piano onset detection experiment (Table 4) and the golf swing event sequencing experiment (Table B.2).

Table B.4: Smoking Puff Detection Ablation Study. Comparison of the deep model trained with $\mathcal{L}_{\text{SLL}}$ and $\mathcal{L}_{\mathcal{S}\text{oftLoc}}$ as loss function with respect to noise distribution $\lfloor \mathcal{N}(0, \sigma^2) \rceil$. We report the mean (standard deviation) of ten 6-fold cross-validated $\boldsymbol{F_1}$**-scores**.

	$\sigma, \delta = 0$	1	2	3	4
$\mathcal{L}_{\text{SLL}}$	93.0 (2.7)	89.7 (3.7)	87.1 (5.0)	82.3 (6.2)	78.3 (7.7)
$\mathcal{L}_{\mathcal{S}\text{OFTLOC}}$	**93.1** (2.5)	**90.6** (3.4)	**87.8** (4.1)	**83.6** (5.2)	**79.0** (6.9)

Table B.3: Smoking Puff Detection. Comparison of LR-M (Adams and Marlin 2017) and the deep model trained with $\mathcal{L}_{\text{CE}}$, $\mathcal{L}_{\text{LS|CE}}$, $\mathcal{L}_{\text{LS|SE}}$, and $\mathcal{L}_{\mathcal{S}\text{oftLoc}}$ with respect to misalignment distributions $\mathcal{B}(-\delta, \delta)$ and $\lfloor \mathcal{SN}(0, \sigma^2, \alpha = -2) \rceil$. Reported metrics are the mean and standard deviation of ten 6-fold cross-validated $\boldsymbol{F_1}$**-scores**.

		$\delta, \sigma = 0$	1	2	3	4
$\underline{\mathcal{B}}$	LR-M	—	65.5 (14.5)	54.9 (20.4)	44.1 (19.7)	51.8 (19.8)
	$\mathcal{L}_{\text{CE}}$	—	41.7 (15.3)	28.3 (14.5)	26.6 (15.3)	22.8 (15.1)
	$\mathcal{L}_{\text{LS|CE}}$	—	60.7 (6.7)	53.0 (8.8)	43.7 (10.1)	34.8 (13.0)
	$\mathcal{L}_{\text{LS|SE}}$	—	45.2 (8.3)	54.7 (9.2)	45.1 (11.6)	35.4 (11.8)
	$\mathcal{L}_{\mathcal{S}\text{oftLoc}}$	—	**90.8** (3.3)	**87.0** (4.7)	**81.7** (7.2)	**72.4** (10.1)
$\underline{\mathcal{SN}}$	LR-M	—	79.7 (10.4)	68.3 (15.6)	61.4 (20.7)	54.7 (18.2)
	$\mathcal{L}_{\text{CE}}$	—	57.6 (16.6)	27.8 (13.7)	20.0 (13.9)	16.1 (14.4)
	$\mathcal{L}_{\text{LS|CE}}$	—	53.8 (9.8)	49.6 (8.5)	43.9 (10.7)	41.1 (8.6)
	$\mathcal{L}_{\text{LS|SE}}$	—	57.0 (8.2)	52.1 (8.4)	48.3 (7.9)	44.4 (9.6)
	$\mathcal{L}_{\mathcal{S}\text{oftLoc}}$	—	**90.4** (3.9)	**88.2** (5.0)	**84.2** (6.1)	**79.1** (9.0)

C Piano Onset Detection Experiment

Extreme Noise Settings Table 3 (in the main text) depicts the strong invariance of our $\mathcal{L}_{\mathcal{S}\text{oftLoc}}$ model to label misalignment on a broad array of noise levels (i.e., up to $\sigma = 200\text{ms}$). In this section, we evaluate the model's performance on an even greater range of noise levels to fully assess its behavior in extreme settings. To that end, additional piano onset detection experiments, with noise levels up to $\sigma = 1000\text{ms}$, were conducted following the protocol described in Section 5.3. The results are displayed in Figure B.3.

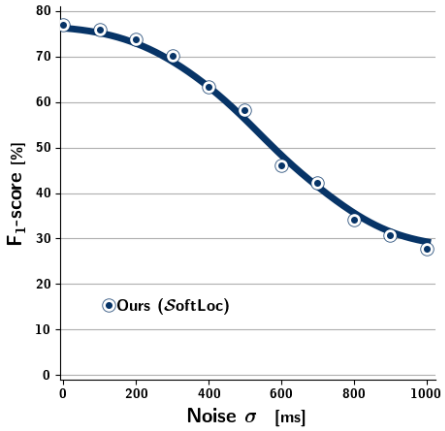
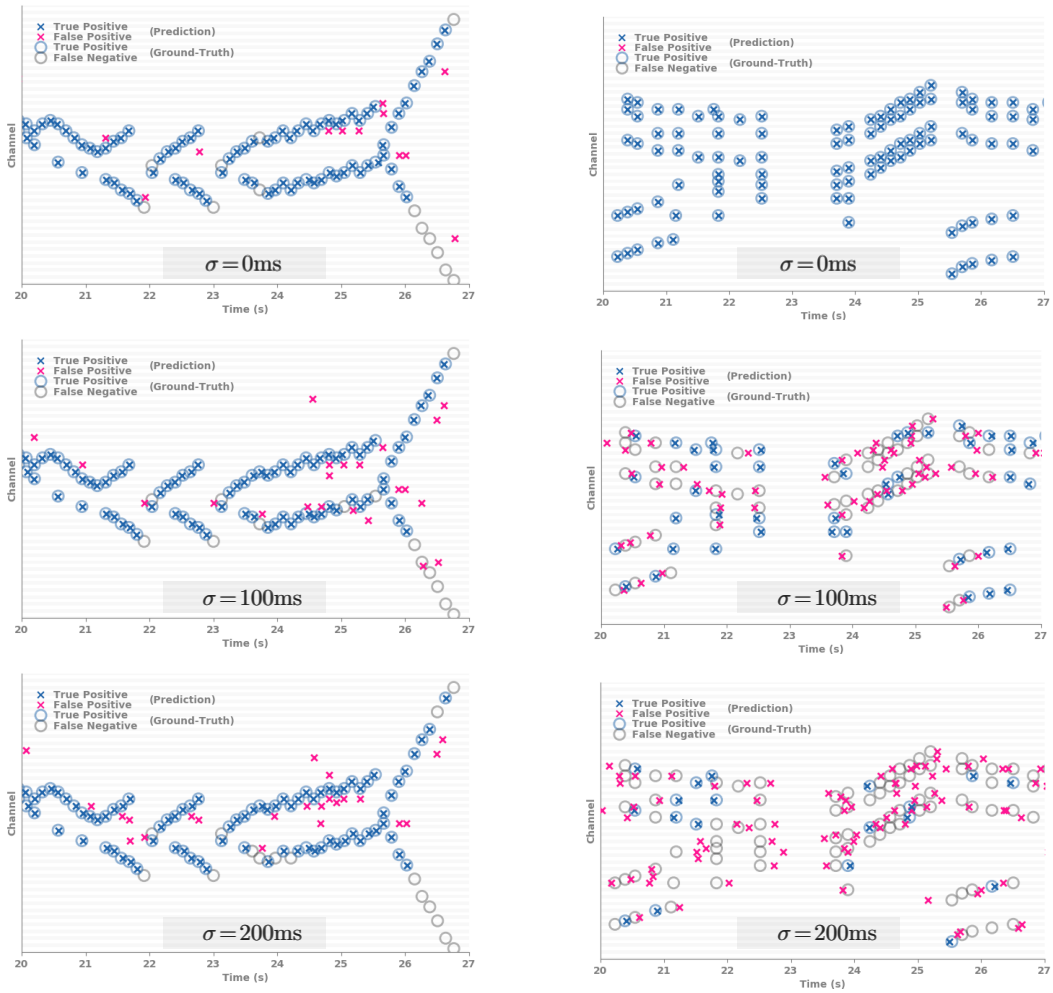


Figure B.3: Piano Onset Detection in Extreme Noise Settings. F_1 piano onset detection performance of our $\mathcal{L}_{\mathcal{S}\text{oftLoc}}$ model ($s_M = 100\text{ms}$) as a function of label misalignment.

Overall, this figure confirms the remarkable robustness of our $\mathcal{L}_{\mathcal{S}\text{oftLoc}}$ model to label misalignment. While the absolute performance unsurprisingly decreases as the training data becomes less accurate, the detection capability of the model in noisy settings outshines any classical approach (see Table 3). Note that a fixed softness value ($s_M = 100ms$) was used across all noise levels and that these results could further be improved by increasing the model softness s_M (see Section 5.4).

Noisy Labels and Ground-Truth Discrepancy To further illustrate the complexity of the localization task when annotations are subject to misalignment, we compare the noisy training labels with the actual ground-truth event locations. Figure B.2(b) displays an example of the quality of the training labels. Obviously, in the noise-free setting (i.e., $\sigma = 0\text{ms}$), the localization is spotless as the training labels and the ground-truths are identical. However, as the noise level increases, the proportion of labels that stay within the 50ms tolerance window decreases significantly. More precisely, the performance (i.e., F_1-score) of the labels themselves is 68.2%, 39.8%, and 23.7% for σ equal to 50ms, 100ms, and 200ms respectively. This contrasts with the performance of our approach, which appears almost invariant to the noise level (see Figure B.2(a)).



(a) Out-of-sample predictions of our $\mathcal{L}_{\mathcal{S}\text{oftLoc}}$ model trained on data subject to various noise levels. *(Schubert—Piano Sonata in A minor, D 784, Opus 143, 3. Mov)*

(b) In-sample performance of the noisy training labels themselves (*as predictions*) when compared to the clean ground-truth. *(Liszt—Hungarian Rhapsody No. 10)*

Figure B.2: Robustness of Predictions to Noisy Labels. Difference between (a) the consistency of the predictions and (b) the quality of provided labels for training across various noise levels σ, ranging from noise-free ($\sigma = 0\text{ms}$) to extremely noisy ($\sigma = 200\text{ms}$).