We would like to thank all the reviewers for the insightful and detailed comments.

Reviewer 1 Each triggering kernel $\phi_{u_i u_j}$ in HP only connects *past* events of type u_j to the CIF, $\lambda_{u_i}(t)$, of type u_i. The CIF by definition is *conditioned* on the history (all types of events that occur before t), so $\lambda_{u_i}(t)$ only depends on the history, not on $\lambda_{u_j}(t)$. The factorization can be derived directly from the general form of the probability density (line 68-71) due to the definition of CIF (line 63) and is not restricted to HPs. Also, we derive the method for one **target** event type, but model the dependencies between that type and *all* types of events through $x(t)$ (line 80-91): we model each $\lambda_u(t), u = 1, \ldots, U$, with a GP, but the input $x(t)$ of the GP depends on all U types of events that occur before t.

A key difference between conditional GP and variational sparse GP [Titsias 2009, Lloyd et al. 2015] is in the flexibility of the models. Inducing points in variational sparse GP are marginalized out in the final inference, so they do not add any flexibility to the model (they only improve computational efficiency). Conditional points in conditional GP are kept in the final inference as being conditioned on (e.g., Eq. 4), so they add flexibility and therefore help the model to store the dependency information learned from the data. We have results comparing conditional GP and variational sparse GP in the supplementary material (Section F).

The dependency between events can have variance in the data: given the same history, the CIF can still vary in different realizations. Noiseless conditional points cannot capture the variance. That is why we introduce noise in f_Z, generalizing the noiseless version. S_ϵ learned from the data captures the variance and covariance of the conditional points. The entries in S_ϵ are small in many cases, but it is nice to allow it to adapt to the data (e.g., to have large entries when the dependency between events shows much variance in the data).

We apologize for the ambiguity. We did not mean CGPRPP cannot model bursty events, but tried to stress that it is easier for HP variants to model them, since the data satisfy their assumption (bias). The majority of the inferior results on IPTV and MIMIC are seemingly caused by overfitting (CGPRPP has better training likelihood), the root cause of which could be change in data distribution between training and test sets.

We note that both HP-GS and HP-LS are *nonparametric* and flexible, which is why we picked them as baselines. Thanks for pointing out other works. However, 1) [Xie et al.] only consider *univariate* event sequences, so their method cannot be applied to *multivariate* event sequences we consider. 2) [Rousseau et al.] focus on theoretical analyses of posterior convergence rates. Although they have a "numerical illustration", they do not evaluate predictive performance nor compare against existing methods. The inference algorithm is only briefly mentioned lacking details. 3) Similar to HP-GS [Xu et al.], in both [Xie et al.] and [Rousseau et al.], only the *triggering kernels* are nonparametric, while the CIF has the same form as HPs, so they also cannot model a mix of excitations and inhibitions. In contrast, CGPRPP can model it, because the CIF is nonparametric, so CGPRPP has the same advantage. 4) None of them compare against HP-GS or HP-LS in the experiments, so their methods are not necessarily more SOTA than HP-GS or HP-LS.

Reviewer 2 We believe you mean Eq. 1 that defines the kernel. As noticed, $x_d(t) = t - s_u^q(t)$ is undefined when the q-th (from last) event of type u does not exist yet at time t (e.g. when there is no type u event in the history). Conceptually, we augment each dimension $x_d(t)$ with a new dimension $x_{D+d}(t)$: $x_{D+d}(t) = 0$ if $x_d(t)$ is undefined and 1 otherwise, increasing the dimensionality of $x(t)$ from D to $2D$. Then instead of leaving $x_d(t)$ undefined, we can assign a special value $x_d(t) = \infty$ (or a very large number), when the q-th (from last) event of type u does not exist. Then define $\mathbb{I}[x_d(t)] = 0$ if $x_d(t) = \infty$ and 1 otherwise, or equivalently define $\mathbb{I}[x_d(t)] = x_{D+d}(t)$. The overall kernel is valid due to the kernel composition rules and that K_1 and K_2 are valid kernels on augmented dimensions and original dimensions. K_2 is widely used. K_1 is a valid kernel because it is an inner product on augmented dimensions.

Reviewer 3 We cut the conclusion text due to the space limit. We plan to work on the text and hope to find the space to add it back and also add an illustration. We will fix the inconsistency in the bibliography (thanks for noticing it).

Reviewer 4 By "focus on one type" we mean deriving the method for one **target** event type, but the dependencies between that type and *all* types of events are modeled through $x(t)$ (line 80-91): we model each $\lambda_u(t), u = 1, \ldots, U$, with a GP, but the input $x(t)$ of the GP depends on all U types of events before t. Based on the factorization of the density (line 68-71), we can repeat the derivation for all target event types (same equations with different target event types). Losses (negative log-likelihood) from different sequences are summed for learning (Eq. in line 79 with $\log$).

In Remark 1, we mean it is isolated from the time *denoted by* t as in $x(t)$, which is the *absolute* time. We will make it clearer in the paper. We note that whether Assumption 1 breaks the mechanics of HP depends on what triggering kernel is used for HP. For example, if the triggering kernel is only nonzero within a bounded interval and zero otherwise, i.e., $\phi(t) = 0$ if $t > A, A \in \mathbb{R}_{>0}$ (interestingly [Donnet et al.] you referenced assume this (Section 1.4) and state it is a "very common" hypothesis (Section 2.1)), then Assumption 1 will not break the mechanics. Thanks for pointing out other works. We will cover them in related work.