We thank the reviewers for their time and effort reviewing our paper, and for the positive appraisal of its content. Our work is the first semi-supervised model to achieve SOTA results in a sparsely-labeled data regime on a wide variety of animal datasets. We firmly believe this work will make a valuable contribution to the NeurIPS community.

To begin, we emphasize a few critical points shared among reviewers:

1) **Methodological novelty.** Our model is a novel semi-supervised learning approach with spatial and temporal cliques for large-scale sparsely-labeled videos. In the field of human pose estimation (HPE), there have been significantly fewer methods in the video domain due to a limited number of large-scale benchmarks for video tracking. Thus the significance of semi-supervised models utilizing unlabeled frames is nonnegligible. Two papers in HPE closely related to our work are Song et al (CVPR2017) and Bertasius et al (NeurIPS2019). The former has the spatial-temporal graph but doesn't consider unlabeled frames, and the latter works for sparsely-labeled data but without any spatial constraint. Moreover, in the field of animal pose estimation (APE), we believe there is no such work yet. Another contribution is formulating the semi-supervised learning in a Bayesian framework. We employed a structured variational inference method for the hybrid Bayesian latent variable model which hasn't been done in this field. We apologize if these points were not clear in the main paper, and we will emphasize them more clearly in the revision.

2) **Lack of comparison to other approaches.** We would like to clarify that we use DLC as the underlying architecture for DGP to show that when such a model is recast in the proposed semi-supervised framework–which considers unlabeled frames and spatial and temporal cliques–the overall model performance improves. We didn't find other semi-supervised approaches in the APE literature that incorporate unlabeled frames, so instead our experiments consisted of ablation studies comparing different components of DGP. Our framework is easily compatible with other architectures; we have already run comparisons of DGP with a DPK (Graving et al elife2019) architecture instead of DLC, and will include these results in the revision.

3) **Missing details.** We apologize for missing some details. We aimed at writing a neuroscience-application focused paper, thus unavoidably moved some modeling and experiment details to the supplementary. We will definitely consider moving them back if it helps with model clarity and polish the experimental details in the revision.

4) **Missing link.** We would like to sincerely apologize for the missing link issue. The submission system wiped off the links and we didn't carefully check them after submission. We will fix this issue in the revised version.

Reviewer 2:

1) **Spatio-temporal models.** The focus of this work is not in the complexity of the spatio-temporal cliques, but rather in how to incorporate these cliques into a hybrid graphical model with latent variables. As mentioned in the paper, the proposed spatio-temporal cliques can include additional dynamics or other structural constraints; this is an important topic for future work. Here we provide a set of cliques which achieve good performance in a wide variety of datasets.

2) **Missing details in sec. 3 and experiments.** We will include more details for the training and testing splits in Fig 3. We split the data in two separate train and test sets. To train our model, we employed a pre-trained Resnet-50 on Imagenet. At test time, in DGP-NN we pass test frames into the trained network and calculate the 2D locations. During test time, we can optionally run one additional E-step optimization for DGP which functions as a post-hoc smoothing without changing the network parameters. We will clarify these points in the revised version.

3) **Related work.** We mainly respond to this point in **Methodological novelty**. The related papers pointed out by the reviewer didn't consider unlabeled frames. This discrepancy is non-trivial from a modeling perspective. However, we thank the reviewer for pointing out these references and will include them in the revised version.

Reviewer 3:

1) **How much do potentials matter.** The comparison between DLC semi and DGP indicates the benefit of the spatial and temporal cliques. But as the reviewer pointed out, we didn't show cross-ablation studies to highlight the impact of the individual cliques. We have already done these analyses and will include them in the revised version. We also thank the reviewer for pointing out the work by Liu et al and we will cite this paper in the revised version. In DGP, the proposed temporal clique can be extended to a more elaborate function e.g. optical flow. As mentioned earlier, our paper's focus is to propose a hybrid latent variable model where information from scarce human-labeled frames can be propagated to unlabeled frames. Our current construction of the spatial and temporal cliques is simple and flexible, but these can be extended to other more complex forms.

2) **Notation issues in the supplement.** We apologize for these issues. These will be corrected in the revised version.

Reviewer 4:

1) **The model is not end-to-end.** This seems to be a misunderstanding: our model is an end-to-end model. DGP combines a graphical model and a neural network in a hybrid model, where the loss function includes the parameters from both of these components. This is different from - and empirically superior to - training a baseline model (such as DLC) and applying a Kalman filter to the network outputs afterwards.

2) **Graph neural network.** The graph neural networks (GNN) literature that we have encountered pertains to networks where the different nodes have graph structures. In constrast, DGP attaches a CRF-like graph to the output of the neural network, as opposed to changing the network structure. We would ask the reviewer for additional clarification if we missed some GNN literature more closely related to our proposed model or if we misunderstood the point being made.