We apply random forces on the contour of the fabric to deform and move it around, where we encode the action as a 6-dimensional vector: the first three is the coordinate of the dragged point, and the other three indicate the movement, which will then be concatenated with the embedding of every keypoint.

**R1:** Why Structural causal models (SCMs)? SCMs are the core of causal modeling & inference; and the underlying generative process of the physical mechanism (often a system of differential equations) is essentially an SCM, which corresponds to a DAG when unrolled as a causal full time graph [Peters et al. (book)]. In this work, we only assume access to visual observations. The ability to recover an SCM that closely resembles the ground truth SCM will allow the model to perform extrapolation and make counterfactual predictions, as have been demonstrated in the paper (Figure 6 & 7).

**R2:** Action carried out in the fabric domain. We apply random forces on the contour of the fabric to deform and move it around, where we encode the action as a 6-dimensional vector: the first three is the coordinate of the dragged point, and the other three indicate the movement, which will then be concatenated with the embedding of every keypoint.

**R3:** Separation between key-point extraction and relational model learning. We have tried to train the modules jointly in an end-to-end framework. However, due to the interplay between many competing losses, we observed degradation in the perception module and degeneracy of the detected keypoints. Still, this is an open question for future work on better architectures for end-to-end training.

We will release code to support reproducibility and also correct the typesetting, citation format and adjust the language to be precise. We will also add discussion in broader impact of V-CDN in camera-ready version.