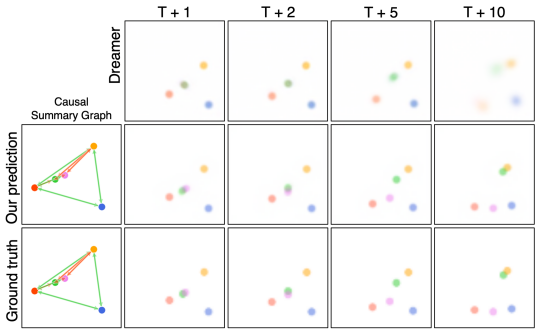
We thank all the reviewers for their constructive comments.

Comparison with baselines (R1, R2, R3).

The primary goal of this paper is to perform the one-shot discovery of structural causal models (SCMs) from observational videos.

Comparing Keypoint Predictions. As R2 pointed out, there are, unfortunately, not many baselines for us to compare. In the paper, we included a baseline that uses a simpler prediction module using graph-nets without causal SCM modeling. V-CDN significantly outperforms the graph-based prediction, indicating the importance of accurate modeling of the causal mechanism (Figure 6 (d) in the main paper and Figure 2 in the supplement).



Comparing Video Predictions. Making predictions directly on a pixel level without the intermediate structures won't be able to recover the SCM, hence lacking the ability to perform counterfactual prediction or extrapolate to unseen graphs. Still, we follow the reviewers' suggestion by including an additional baseline that predicts directly over the pixels. We compare with a state-of-the-art model-based reinforcement learning method, *Dreamer*[Hafner et al ICLR 2020], which uses an encoder-decoder model to learn latent-space dynamics model from image inputs and reconstruct future image observations. For a direct comparison, we train a visualization module that maps the predicted future keypoints from V-CDN back to the pixel space.

The above figure shows the results. *Dreamer*'s prediction deviates from the ground truth and quickly becomes blurry, suggesting the importance of the structured intermediate representation used in our model.

Selection of the environments (R1, R4). The objective of this paper is to propose a causal discovery method for time-series data with image observations. This is a hard problem, especially with the one-shot test-time prediction setup, which has few baselines. The method is in no way specific to these domains and the domains are chosen in order to evaluate the generalizability of the method. The particle domain has a observational state in direct correspondence with the ground truth node variables in the DAG, which allows us to perform a systematic analysis of the performance across varying latent true generative models. While the fabric domain exhibits causal discovery wherein causal variables do not trivially correspond to features in images; hence a single reduced-order model over keypoints can model different types of fabrics. We are not aware of any other learning-based method that can predict the fabric state and its evolution without modeling techniques that are specific to specific fabric shapes and classes. In contrast, our framework can handle the variability of fabric structure and generalize to new shape variations and fabric parameters.

Difficulty of the task (R1, R4). These problems are particularly challenging because, at test time, the model has to extrapolate to unseen edge parameter ranges (Fig. 7), edge types, and graphs of different sizes from training (Fig. 6). Baselines, even with graph-structured prediction models, cannot cope with such out of distribution generalization.

Applicability of the proposed method (R4, R1). It may be a misunderstanding that our method is specific to the problem of "tracking masses that are possibly connected by rigid links or springs" (R4). As shown in the Fabric environment, our method can work with complicated deformable objects, where there are no explicit "masses" for our model to track. Instead, our model extracts a structured representation and reason about the dependency structure directly from fabric images. Meanwhile, our method does not assume to know the specific physical equations that describe the interactions within the environment. Instead, we use graph neural networks to learn the effect of the interactions from data and have shown in the Fabric domain that our model can model the combination of bending and stretching force within the cloth, where writing down the analytical physical equations may be very hard as we are operating in a reduced-order representation.

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.