# Climbing the Ladder of Causation with Counterfactual World Modeling

**Rahul Venkatesh*  Honglin Chen*  Klemen Kotar**
**Kevin Feigelis  Wanhee Lee  Daniel Bear  Daniel Yamins**
Stanford University
* denotes equal contribution

## Abstract

**While language models have begun to show signs of understanding causal relationships, vision models seem to lag behind. We introduce Counterfactual World Modeling (CWM) — a visual world model trained for future prediction that demonstrates capabilities analogous to various levels of Pearl's "Ladder of Causation"[1]. A key finding of this paper is that mid-level vision structures can be formulated as counterfactual queries to CWM, enabling their extraction under a unified, self-supervised architecture. This not only moves closer to a human-like learning process, but also reduces the reliance on expensive annotated datasets for training task-specific models – a long-standing predicament in computer vision.**



**Keywords: Mid-level vision, counterfactuals**


## Introduction

Despite advances in modeling high-level cognitive functions like language understanding, algorithms lag behind human-like visual understanding. They have two shortcomings: 1) Models are not task-general – e.g. a segmentation model can't be used for depth estimation and training requires costly annotated datasets (Kirillov et al., 2023). This is unlike humans, where useful structures are extracted from a unified model (Hong & Yamins, 2016). 2) Vision models struggle to answer causal questions – e.g. object movement due to external force (Bear et al., 2021). In Pearl's "ladder of causation", current computer vision models are arguably closer to the lower rungs (Pearl & Mackenzie, 2018), while language models exhibit signs of more advanced abilities (Li, Yu, & Ettinger, 2022). We now show how CWM seeks to address these challenges by climbing Pearl's ladder of causation.

[1] Pearl's causal hierarchy has three levels of increasing complexity: "Association" involves identifying patterns; "Intervention" entails predicting outcomes as a result of changes to stimuli; and "Counterfactuals" deals with reasoning about hypothetical scenarios.

### Level 1: Learning associations via temporally-factored masked prediction

We train a transformer architecture that reconstructs masked observations of video frames to learn "associations" between spatiotemporal patches of observed video inputs. A subset of the input patches is masked, and the predictor minimizes the mean squared error (MSE) between the reconstructed and the original masked patches. Given the input frame pair, $x_1, x_2 \in \mathbb{R}^{3\times H\times W}$ we train a predictor $\Psi(x_1^{\alpha}, x_2^{\beta}) = \hat{x}_2$ which receives the first frame $x_1$ and the second frame $x_2$ with masking ratio $\alpha, \beta \in [0,1]$ and reconstructs the masked patches of $x_2$ (see Figure 1a). We set $\alpha$ to 0 and $\beta$ to 0.90 – a core hypothesis of our work is that such an asymmetric masking policy, with high masking in the second frame, makes the predictor $\Psi$ learn to concentrate scene transformations into the embeddings of a few visible patches in $x_2$. This enables meaningful control over the predictions via patch-level modifications. As we will demonstrate next, this choice of masking policy is what enables the model to make a leap towards the higher rungs of the causal heirarchy. CWM is trained on Kinetics – a dataset comprising of YouTube videos (Kay et al., 2017).

### Level 2: Interventions via patch-level prompting

With a pre-trained predictor, we can apply interventions to input stimuli by modifying the patches in $x_2$ that control scene transformations. These interventions are akin to the *do* operator introduced in Pearl's framework and are considered more powerful than "Associations" as they involve making predictions about novel stimuli that are possibly outside the training distribution (Goldberg, 2019). To formalize the procedure of intervention, we first define a prompt $p$ as a set of video frames input to the predictor:

$$p = \{x_1, x_2 \,|\, x_1, x_2 \in \mathbb{R}^{3\times H\times W}\} \tag{1}$$

An intervention $\bar{p}$ is defined as an input to the predictor that has been modified from the initial prompt $p$ to change the outcome of the predictor (See Figure 1b).

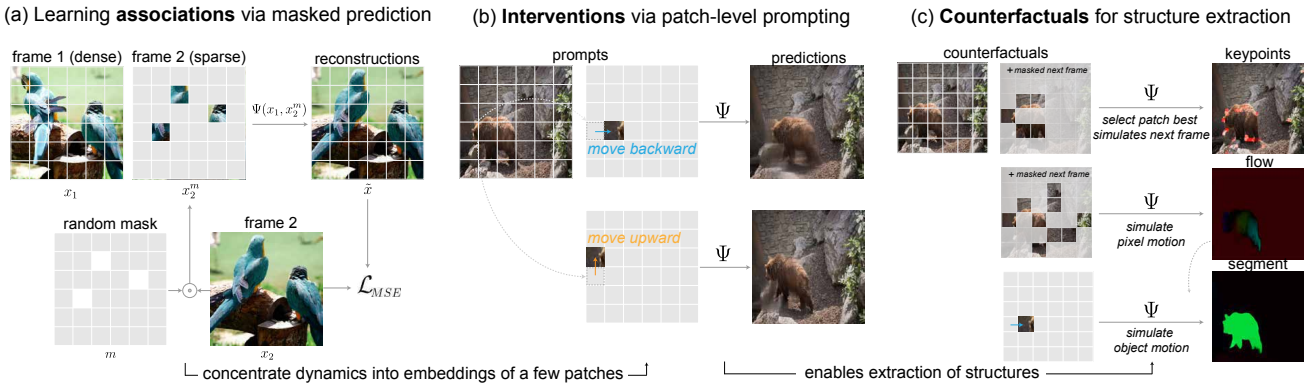



Figure 1: **Climbing the Ladder of Causation with the CWM framework: (a) Learning associations via temporally-factored masked prediction.** Given a frame pair input, the predictor takes in dense visible patches from the first frame and only a sparse subset of patches from the second frame as inputs, and learns to predict the masked patches. This policy encourages the model to concentrate scene dynamics into embeddings of a few patches. **(b) Interventions via patch-level prompting.** As a result of the temporally-factored masking, we can intervene by modifying one or a few visual patches in the prompt and steer the outcome of the predictor. **(c) Counterfactuals for structure extraction.** Multiple vision structures can be extracted by comparing the results of an intervention to alternative futures (e.g. observed ground truth or observed predictions).

### Level 3: Counterfactuals for structure extraction

Next, we discuss how mid-level vision structures can be specified as counterfactuals (Pearl & Mackenzie, 2018). These are more powerful than interventions as they involve comparisons between the current observed prediction $\Psi(p)$ or observed ground truth $x$ and retrospective intervention outcomes, $\Psi(\bar{p})$ (see Figure 1c). We describe these specifications below:

**Keypoints** CWM provides a category-agnostic definition of keypoints as patch locations in $x_2$ that, when revealed to the predictor, yield the lowest reconstruction error as defined by the loss function, $\mathcal{L}$. The set of keypoints on $x_2$ is defined as:

$$K(x_1, x_2, n) = \underset{k \subset I, |k|=n}{\arg\min}\ \mathcal{L}(\Psi(p), \Psi(\bar{p}))$$
$$\text{where} \quad \bar{p} = \{x_1, x_2^m \,|\, x_2^m \text{ is visible at } k\} \tag{2}$$

Here, $I$ refers to the complete set of patch locations of an image, and the intervention $\bar{p}$ is the modification of the original input $p = \{x_1, x_2\}$, where the second frame $x_2^m$ is masked everywhere except at keypoint locations.

**Optical flow** is the task of estimating per-pixel motion between video frames (Teed & Deng, 2020). To compute this, we introduce an intervention that adds a small perturbation to the pixel in the first frame and estimate pixel motion by localizing the perturbation response in $\Psi(\bar{p})$. Given a prompt $p = \{x_1, x_2^\beta\}$ and a location $(i, j)$, we construct an intervention $\bar{p} = \{x_1 + \delta_{ij}, x_2^\beta\}$, which adds a small perturbation $\delta_{ij}$ to the first frame at the pixel location. With a perturbed first frame, the predictor propagates the perturbation in the next frame, under the original scene transformations specified by $x_2^\beta$. The corresponding pixel location in the next frame can be localized by finding the peak of the perturbation response. The flow at $(i, j)$ can then be defined using the following equation:

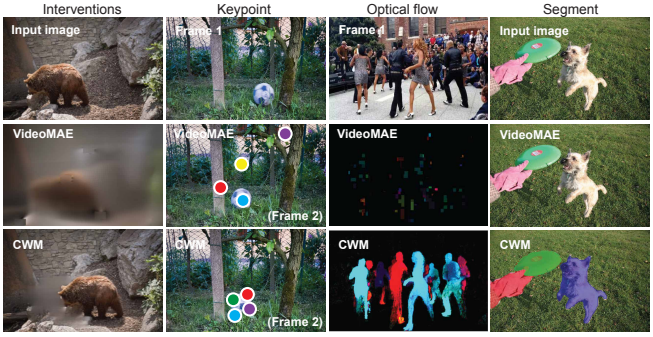
$$F_{i,j}(x_1, x_2) = \underset{I}{\arg\max}\ |\Psi(\bar{p}) - \Psi(p)| - (i, j) \tag{3}$$

**Segmentation** is defined as a grouping of stuff that moves together under physical actions (Spelke, 1990). CWM extracts segmentation by motion interventions which simulate object motion at a pixel location, followed by grouping parts of the image that move together. Given an image $x$ as input, we define an intervention $\bar{p} = \{x, \bar{x}^m\}$, where $\bar{x}^m$ is produced by revealing only a few patches in $x$ and translating them by a small offset. With a temporally-factored masked predictor, moving a few patches in the prompt will cause the entire object to move in the resultant intervention outcome $\Psi(\bar{p})$. Segments can then be extracted by thresholding the flow between $\Psi(\bar{p})$ and the input image:

$$S(x) = F(x, \Psi(\bar{p})) > 0 \tag{4}$$

Multiple objects can be discovered by iteratively extracting segments at locations that are not part of a discovered object.

(a) Intervention outcomes and extracted vision structures.




(b) Quantitative comparisons.

| Methods | Interventions (FID ↓) | Flow (F1 ↓) |
|---|---|---|
| VideoMAE (tube-masking) | 213.4 | 56.3 |
| CWM (temporally-factored masking) | **25.4** | **46.8** |

Figure 2: **Analysis of intervention outcomes and extracted structures.** Visualizations in Figure a) suggest that it is crucial to use the proposed temporally-factored masking policy during training. VideoMAE trained with tube-masking has notably worse structure extractions – segmentation is particularly poor with no meaningful segments discovered. Table b) provides further evidence for this in terms of the quality of intervention outcomes and optical flow.

## Results

We find that CWM extracts meaningful vision structures using the procedures described above (see Figure 2a). As hypothesized previously, a key aspect of our framework that leads to this ability is the temporally-factored masking policy used for learning associations during training. To test this, we evaluate VideoMAE (Tong & Song, 2022), a closely related video transformer architecture that is trained with an alternate masking policy called tube-masking – known to be effective for learning video representations and achieves state-of-the-art results on tasks like activity recognition. However, we find that when we apply our structure extraction procedures to VideoMAE, it yields poor intervention outcomes and leads to inferior counterfactual queries (see Figure 2a). Additionally, the quality of the intervention outcomes as measured by Frechet Inception Distance (FID) (Heusel & Ramsauer, 2017) on the DAVIS dataset and F1 scores (Geiger & Lenz, 2013) on the SPRING optical flow benchmark (Mehl & Schmalfuss, 2023) reported in Figure 2b also speaks to the importance of using a temporally-factored masking policy during training.

## Conclusion

In this paper we present a simple recipe for building vision models that climb Pearl's ladder of causation. Further, we establish that a practically useful consequence of the climb is that it allows for the extraction of mid-level vision structures in a self-supervised manner from a unified architecture, moving closer to human-like visual scene understanding. Currently, our study is limited to mid-level vision – extending the CWM framework to more complex cognitive tasks like vision and language interaction poses an interesting future direction.

## References


Bear, D. M., Wang, E., Mrowca, D., Binder, F. J., Tung, H.-Y. F., Pramod, R., ... others (2021). Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*.

Geiger, A., & Lenz, P. (2013). Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, *32*(11), 1231–1237.

Goldberg, L. R. (2019). *The book of why: The new science of cause and effect: by judea pearl and dana mackenzie, basic books (2018). isbn: 978-0465097609.* Taylor & Francis.

Heusel, M., & Ramsauer, B. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, *30*.

Hong, H., & Yamins, D. L. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature neuroscience*, *19*(4), 613–622.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., ... others (2017). The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... others (2023). Segment anything. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 4015–4026).

Li, J., Yu, L., & Ettinger, A. (2022). Counterfactual reasoning: Do language models need world knowledge for causal inference? In *Neurips 2022 workshop on neuro causal and symbolic ai (ncsi).*

Mehl, L., & Schmalfuss, J. (2023). Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. *arXiv preprint arXiv:2303.01943*.

Pearl, J., & Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic books.

Spelke, E. S. (1990). Principles of object perception. *Cognitive science*, *14*(1), 29–56.

Teed, Z., & Deng, J. (2020). Raft: Recurrent all-pairs field transforms for optical flow. In *Computer vision–eccv 2020: 16th european conference, glasgow, uk, august 23–28, 2020, proceedings, part ii 16* (pp. 402–419).

Tong, Z., & Song, Y. (2022). Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, *35*, 10078–10093.