Do-Causal-JEPA: Learning Interventional World Models with Graph Surgery Estimator

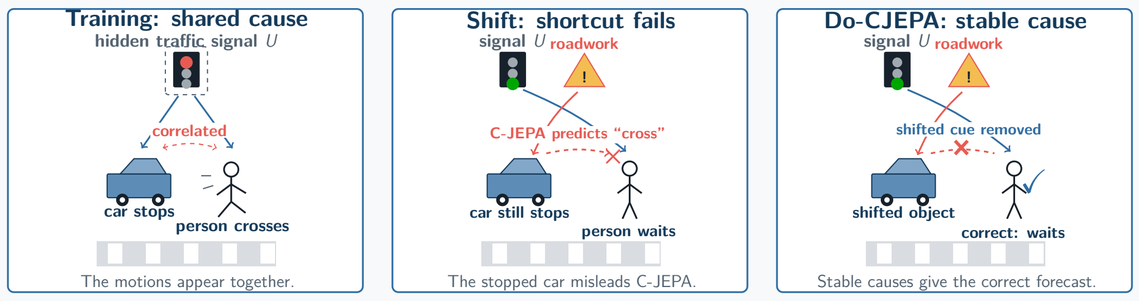
Kenneth Lee[1], Heejeong Nam[2]
[1]Purdue University ▪ [2]Brown University · lee4094@purdue.edu

Take-home message

Motivation: models that predict what happens next can learn shortcuts from training videos and fail when the environment changes, especially when some causes are hidden.

Contributions: we introduce Do-CJEPA, a training method that uses known cause-and-effect relationships to favor predictions that remain stable after a change. Across five repeated Pong experiments, a small amount of this guidance lowered average error, while stronger guidance hurt performance.

Can you predict the next state of an object under mechanism shift with hidden confounding?



1. How C-JEPA works—and what Do-CJEPA changes

C-JEPA: learn a conditional masked predictor

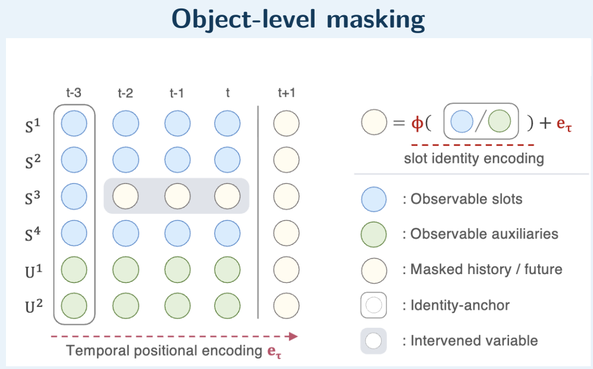
1. Encode. A frozen SAVi encoder produces role-aligned object slots $Z_t^{B,L,R} \in \mathbb{R}^{64}$.

2. Mask whole objects across time. Sample object identities, preserve their first-frame slots as identity anchors, and replace the selected history trajectories and every future slot with learned mask tokens. Optional auxiliary variables remain observable.

3. Predict jointly. A six-layer bidirectional transformer uses the visible objects, anchors, slot-identity encodings, and temporal encodings to recover every masked position.

4. Match observed targets. The loss balances masked-history reconstruction and future prediction.

Object-level masking



A sampled object trajectory is hidden across history while its initial slot anchors identity; all object slots at the future step are masked.

$$\mathcal{L}_{\text{CJEPA}} = \frac{1}{|\mathcal{T}_{\text{hist}}|} \sum_{Y \in \mathcal{T}_{\text{hist}}} \|\widehat{Y} - Y\|_2^2 + \frac{1}{|\mathcal{T}_{\text{future}}|} \sum_{Y \in \mathcal{T}_{\text{future}}} \|\widehat{Y} - Y\|_2^2.$$

Do-CJEPA: add an interventional target for eligible tokens

1. Compile a graph-surgery query. For each masked token $Y = Z_t^r$, FIND-MACS returns $M = \{Y\}$ in the Pong ADMG. With no same-time directed effects,

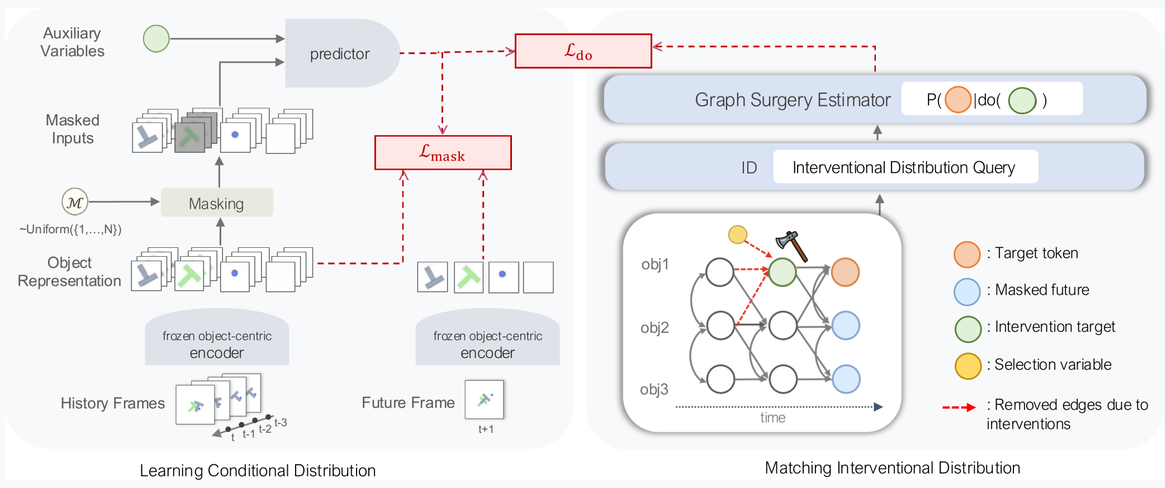
$$X = \text{pa}_G(Y) = \{Z_{t-1}^B, Z_{t-1}^L, Z_{t-1}^R\}.$$

Thus the identifiable query becomes $P(Y \mid \text{do}(X)) = P(Y \mid X)$; a direct child of S is skipped because its mechanism is allowed to change.

2. Learn the interventional mean. One source-trained, role-conditioned ID-GEN model per seed represents $p(Y \mid X)$ as an eight-component diagonal Gaussian mixture. Its analytic mean $\widehat{\mu}_{\text{IDGEN}}(Y; x) = \Sigma_{k=1}^{8} \pi_k(x, r)\mu_k(x, r)$ is cached for every eligible source token.

3. Match both targets. Do-CJEPA keeps the ordinary C-JEPA loss and additionally pulls its prediction toward the stop-gradient interventional mean:

$$\mathcal{L}_{\text{do}} = \frac{1}{|\mathcal{T}_{\text{ID}}|} \sum_{Y \in \mathcal{T}_{\text{ID}}} \|\widehat{Y} - \text{sg}(\widehat{\mu}_{\text{IDGEN}}(Y; x))\|_2^2, \quad \mathcal{L} = \mathcal{L}_{\text{CJEPA}} + \lambda_{\text{do}}\mathcal{L}_{\text{do}}.$$



2. Interventional Pong benchmark

- $T = 16$ frames, three matched roles: ball, left paddle, right paddle; slot dimension 64.
- A one-time soft intervention changes right-paddle position at $t^* \sim \text{Uniform}\{1, \ldots, 15\}$:

$$Z_{t^*}^R \leftarrow \text{clip}\big(0.15\, Z_{t^*}^R + 0.85\,[2\,\text{Beta}(2, 8) - 1]\big)\,.$$

- World models and ID-GEN train and validate on environment-0 source slots only; no shifted environment-1 slots enter optimization. The split contains 10,000/1,000/1,000 initial-condition pairs.
- **Different evaluations:** one-step test: true $Z_{0:14} \to Z_{15}$, possibly with post-shift context. Rollout: true $Z_{0:1}$, then recursive predictions for $Z_{2:15}$.

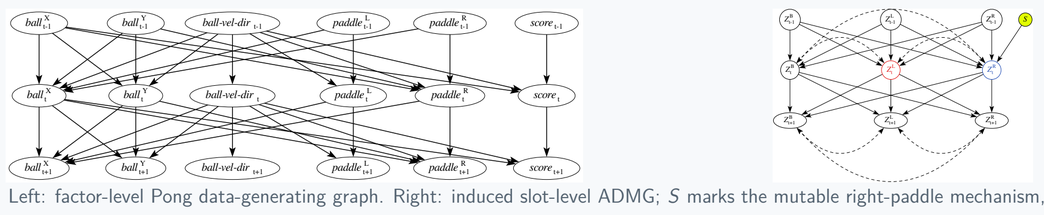
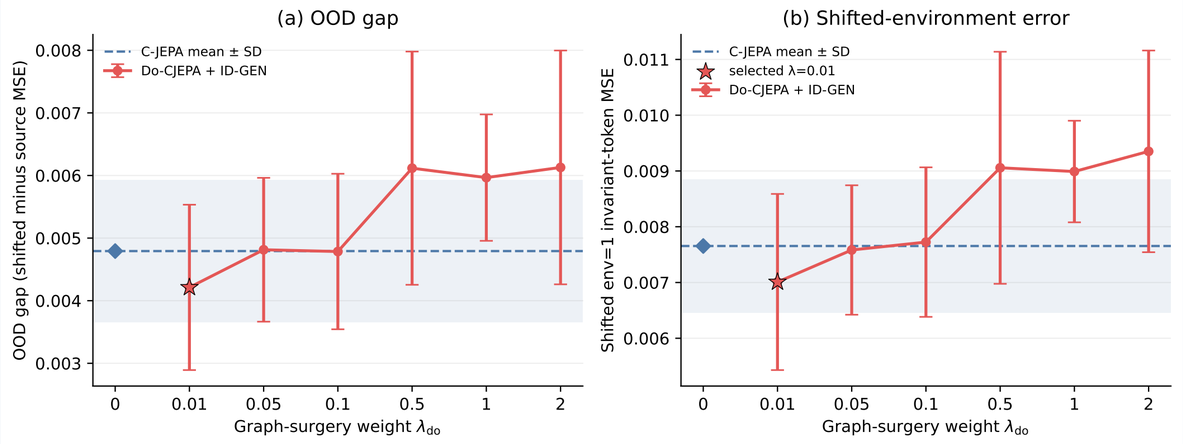


Figure 1. Left: factor-level Pong data-generating graph. Right: induced slot-level ADMG; S marks the mutable right-paddle mechanism, red is the prediction target, and blue is the shifted token.

3. Evaluation and training protocol

- At $t = 15$, invariant-token MSE excludes Z_{15}^R only for the 70/1,000 test pairs with $t^* = 15$. For $t^* < 15$, the shift is already in the observed context and all three roles are scored.
- The OOD gap is $\text{MSE}_1^{\text{inv}} - \text{MSE}_0^{\text{inv}}$; lower is better. All values are mean ± sample SD over five seeds.
- The rollout uses all roles from $t = 2$ onward and is separate from the one-step invariance metric.
- Each role-conditioned ID-GEN estimator and world model is trained for 30 epochs; every sweep point uses seeds $0, \ldots, 4$.

4. Quantitative Evaluation



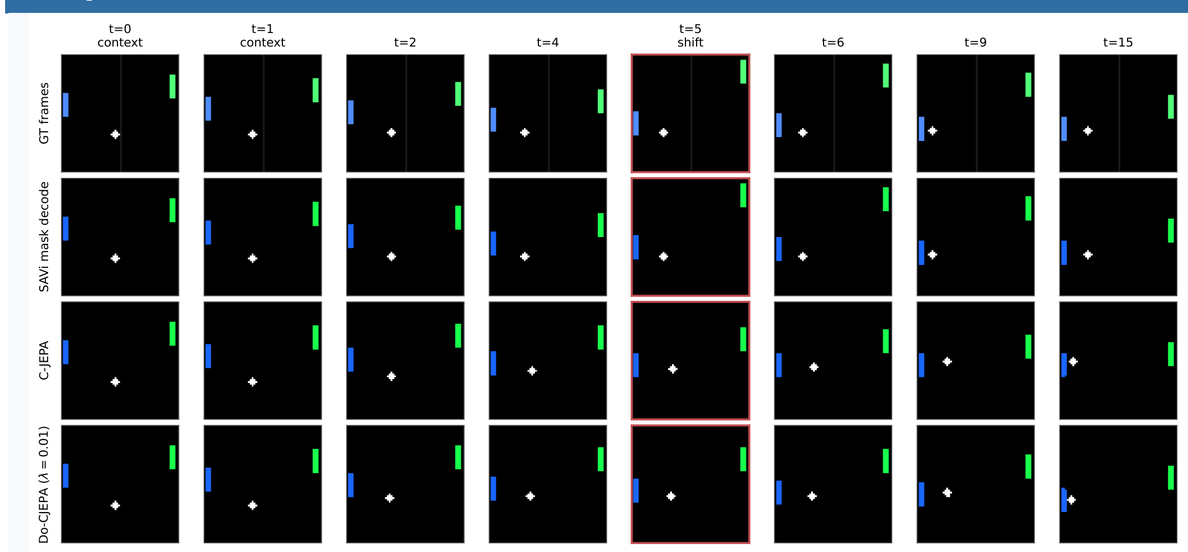
−8.4%
shifted MSE at $\lambda_{\text{do}} = 0.01$

−12.1%
OOD gap at $\lambda_{\text{do}} = 0.01$

Model	Source MSE	Shifted MSE	OOD gap
C-JEPA	0.002862 ± 0.000414	0.007654 ± 0.001195	0.004793 ± 0.001136
Do-CJEPA, $\lambda = .01$	0.002797 ± 0.000292	$\mathbf{0.007009} \pm 0.001580$	$\mathbf{0.004213} \pm 0.001321$

Mean ± sample SD. Paired shifted-MSE and OOD-gap differences favor Do-CJEPA in 4 of 5 seeds. The response is non-monotonic: $\lambda_{\text{do}} \geq 0.5$ degrades both metrics.

5. Qualitative shifted rollout



Top to bottom: ground truth, SAVi-mask rendering, C-JEPA, and Do-CJEPA. Ground truth is supplied only at $t = 0, 1$; from $t = 2$ onward, each one-step prediction is fed back as context. No later ground-truth slots are exposed. The red border marks the unannounced shift at $t = 5$.

6. Insights from Experiments

- A **small** graph-surgery weight can improve mean shifted MSE and OOD gap in this controlled benchmark.
- The selected model is also lower on each shifted role: ball 0.00810 vs. 0.00823, left paddle 0.00163 vs. 0.00170, and right paddle 0.01749 vs. 0.01924.
- The experiment demonstrates a concrete bridge from FIND-MACS identification to learned interventional targets for object-centric prediction.

Interpretation boundary. The $\lambda_{\text{do}} = 0.01$ setting was selected descriptively on the shifted test sweep; no formal significance test is claimed. The Pong graph exercises only the singleton-MACS conditional-generator base case, not general recursive ID-GEN.

Assumptions required by the current pipeline

Known ADMG. The object-level graph and mutable right-paddle mechanism are supplied, not discovered.

Aligned roles. Slots are matched to ball, left paddle, and right paddle before any graph query is formed.

Identifiable target. The extra loss is applied only when graph surgery identifies the interventional distribution.

Temporal graph. With no same-time directed effects, each target uses all three previous-frame roles as parents.

7. Limitations and future work

- Do-CJEPA requires a complete causal graph. Causal discovery with known temporal order should be used to learn the graph needed for invariant next-state prediction.
- Do-CJEPA assumes that encoded slots provide semantically meaningful object representations, with persistent roles that can be matched consistently across a trajectory.
- Graph-surgery regularization applies only when the target interventional query is identifiable from the observational distribution under the assumed graph. Future work will strengthen identifiability results under partial causal graphs.

C-JEPA versus Do-CJEPA at a glance

Aspect	C-JEPA	Do-CJEPA
Training target	Observed masked slot Y	Observed Y plus the cached interventional mean $\hat{\mu}_{\text{IDGEN}}(Y; x)$
Graph usage	No causal graph is used	Known ADMG → Finding intervention set wrt target → identifiable source-data estimator
Eligible tokens	Every sampled masked-history and future target	Extra loss only for identifiable targets that are not direct children of the selection variable
Inference	Six-layer masked transformer	The same transformer; graph surgery and ID-GEN provide training supervision and are not auxiliary inputs at test time

References

- H. Nam, Q. Le Lidec, L. Maes, Y. LeCun, and R. Balestriero. *Causal-JEPA: Learning World Models through Object-Level Latent Masking*. arXiv preprint arXiv:2602.11389, 2026.
- I. Shpitser and J. Pearl. *Complete Identification Methods for the Causal Hierarchy*. Journal of Machine Learning Research, 9:1941–1979, 2008.
- M. M. Rahman, M. Jordan, and M. Kocaoglu. *Conditional Generative Models Are Sufficient to Sample from Any Causal Effect Estimand*. Advances in Neural Information Processing Systems, 37:77269–77315, 2024.