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

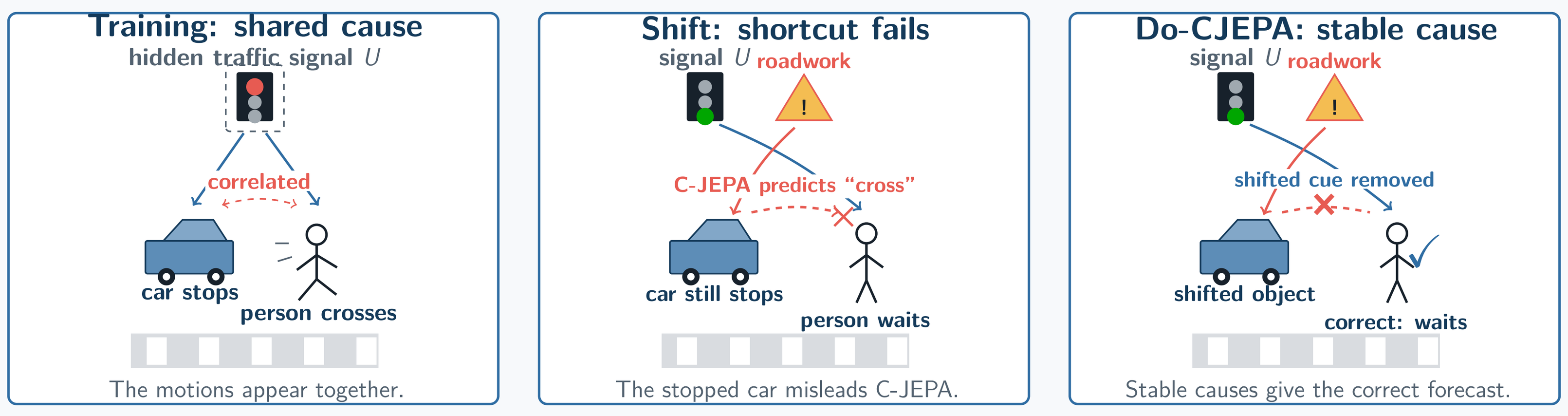
Kenneth Lee[1], Heejeong Nam[2]
[1]Purdue University ▪ [2]Brown University · lee4094@purdue.edu

PURDUE UNIVERSITY BROWN

Take-home message

We introduce Do-CJEPA, which matches changes in model predictions to estimated changes in graph-identified interventional means. In the Interventional Pong experiment, Do-CJEPA reduces online one-step prediction error under mechanism shifts. A gated causal-residual variant combined with source-only four-step rollout training reduces open-loop rollout error relative to both Do-CJEPA and C-JEPA. This comparison evaluates the combined method and does not isolate the effect of gating alone.

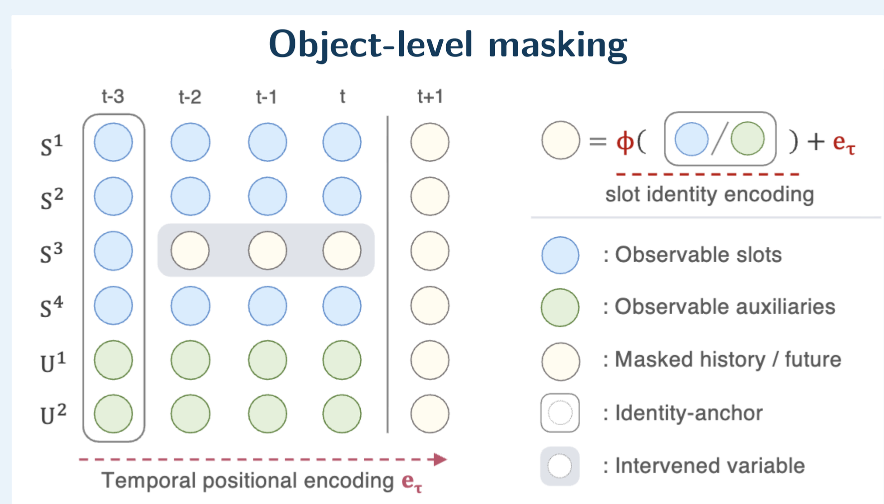
Can you predict the next state of an object under mechanism shift with hidden confounding?



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.



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

$$\mathcal{L}_{\mathrm{CJEPA}} = \frac{1}{|\mathcal{T}|} \sum_{(\tau,r)\in\mathcal{T}} \left\| \widehat{Z}_{\tau}^{r,\mathrm{cjepa}} - Z_{\tau}^{r} \right\|_2^2 .$$

where $\mathcal{T}$ contains the masked batch targets indexed by time τ and object r. The predictor $f_\theta(\bar{Z})_{\tau,r}$ estimates each target from masked sequence $\bar{Z}$, and the loss averages its latent squared error against the observed slot Z_τ^r.

Do-CJEPA: match the interventional response for identifiable targets

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,

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

Thus the identifiable query becomes $P(Y \mid \mathrm{do}(X)) = P(Y \mid X)$; 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 $\widehat{\mu}_{\mathrm{IDGEN}}(Y; x) = \Sigma_{k=1}^{8} \pi_k(x,r)\mu_k(x,r)$ is cached for every eligible source token.
3. Match the interventional response. Evaluate ID-GEN and the predictor at observed parents i and source-donor parents $\tilde{i}$. Do-CJEPA matches their changes while retaining the ordinary C-JEPA loss:

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

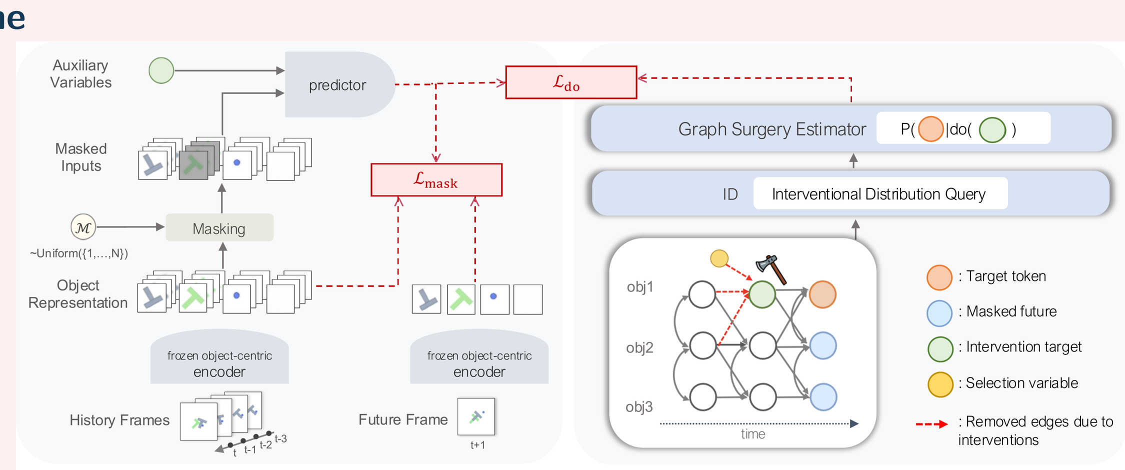
Gated Do-CJEPA: add a gated causal correction

$$\widehat{Y} = \widehat{Y}_{\mathrm{C}} + g_\theta r_\theta, \qquad g_\theta \in (0,1).$$

$$\mathcal{L}_{\mathrm{do}}^{\mathrm{gate}} = \frac{1}{|\mathcal{T}_{\mathrm{ID}}|} \sum_{Y\in\mathcal{T}_{\mathrm{ID}}} \left\| [(g_\theta r_\theta)^{\tilde{i}} - g_\theta r_\theta] - \mathrm{sg}(\widehat{\mu}_Y(\tilde{i}) - \widehat{\mu}_Y(i)) \right\|_2^2 . \qquad \mathcal{L}_{\mathrm{Gated}} = \mathcal{L}_{\mathrm{CJEPA}} + \lambda_{\mathrm{do}}\mathcal{L}_{\mathrm{do}}^{\mathrm{gate}}.$$

It uses the same interventional-response target as $\mathcal{L}_{\mathrm{do}}$ above, but applies it to the change in the gated correction rather than the full prediction. Thus $\widehat{Y}_{\mathrm{C}}$ follows ordinary C-JEPA, while a near-zero gate suppresses each causal correction until it is supported. Here g_θ is a learned token-wise gate, and r_θ is a learned causal residual that proposes how $\widehat{Y}_{\mathrm{C}}$ should change when graph-identified parent values change.

Training pipeline



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 $t^* \sim \mathrm{Uniform}\{1,\ldots,15\}$:

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

- 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 $Z_{2:15}$.

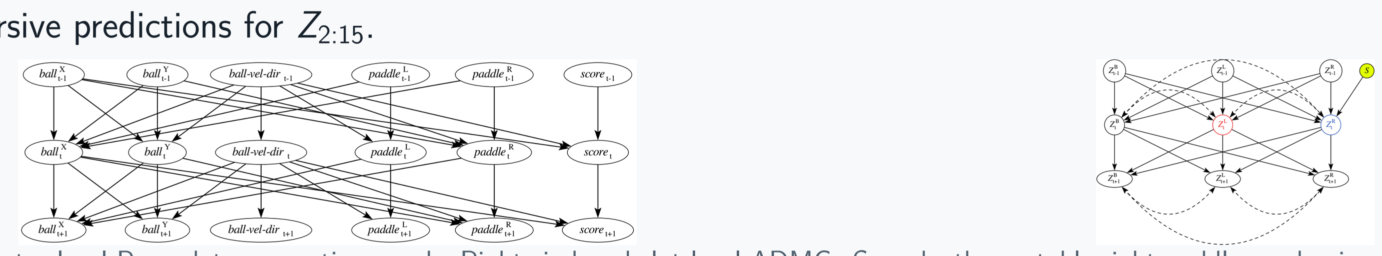


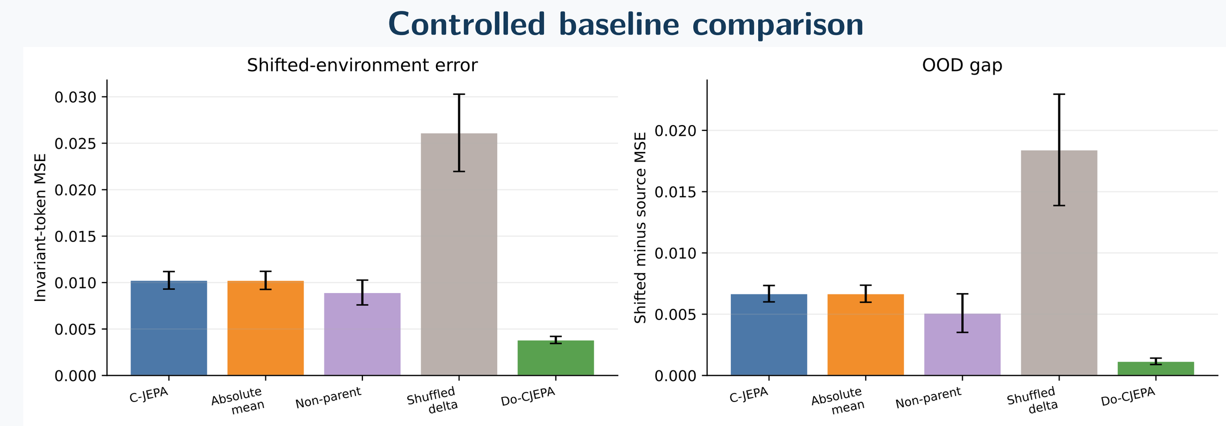
Figure 1. Left: factor-level Pong data-generating graph. Right: induced slot-level ADMG; S marks the mutable right-paddle mechanism, red is the prediction target, and blue is the shifted token.

3. Evaluation and training procedure

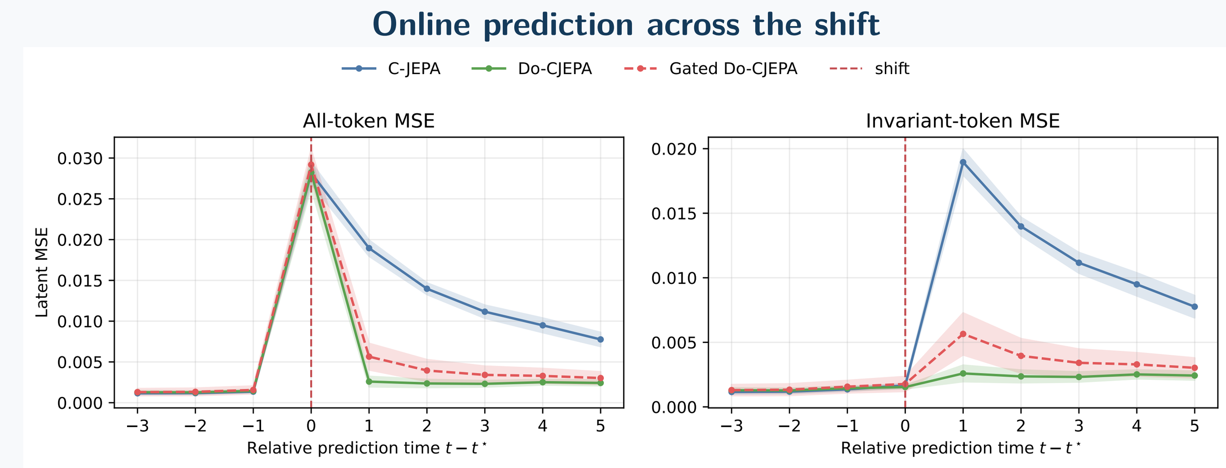
- At $t = 15$, invariant-token MSE excludes Z_{15}^R only for the 70/1,000 test pairs with $t^* = 15$. For $t^* < 15$, the shift is already in the observed context and all three roles are scored.
- The OOD gap is $\mathrm{MSE}_1^{\mathrm{inv}} - \mathrm{MSE}_0^{\mathrm{inv}}$; lower is better. All values are mean $\pm$ sample SD over five seeds.
- Figure 4 evaluates online prediction from observed histories. Figures 5–6 evaluate recursive rollout from only $Z_{0:1}$; all later predictions are fed back as context.
- Each role-conditioned ID-GEN estimator and world model is trained for 30 epochs; each comparison is repeated for 5 seeds.

4. Quantitative evaluation

Controlled baseline comparison



Online prediction across the shift

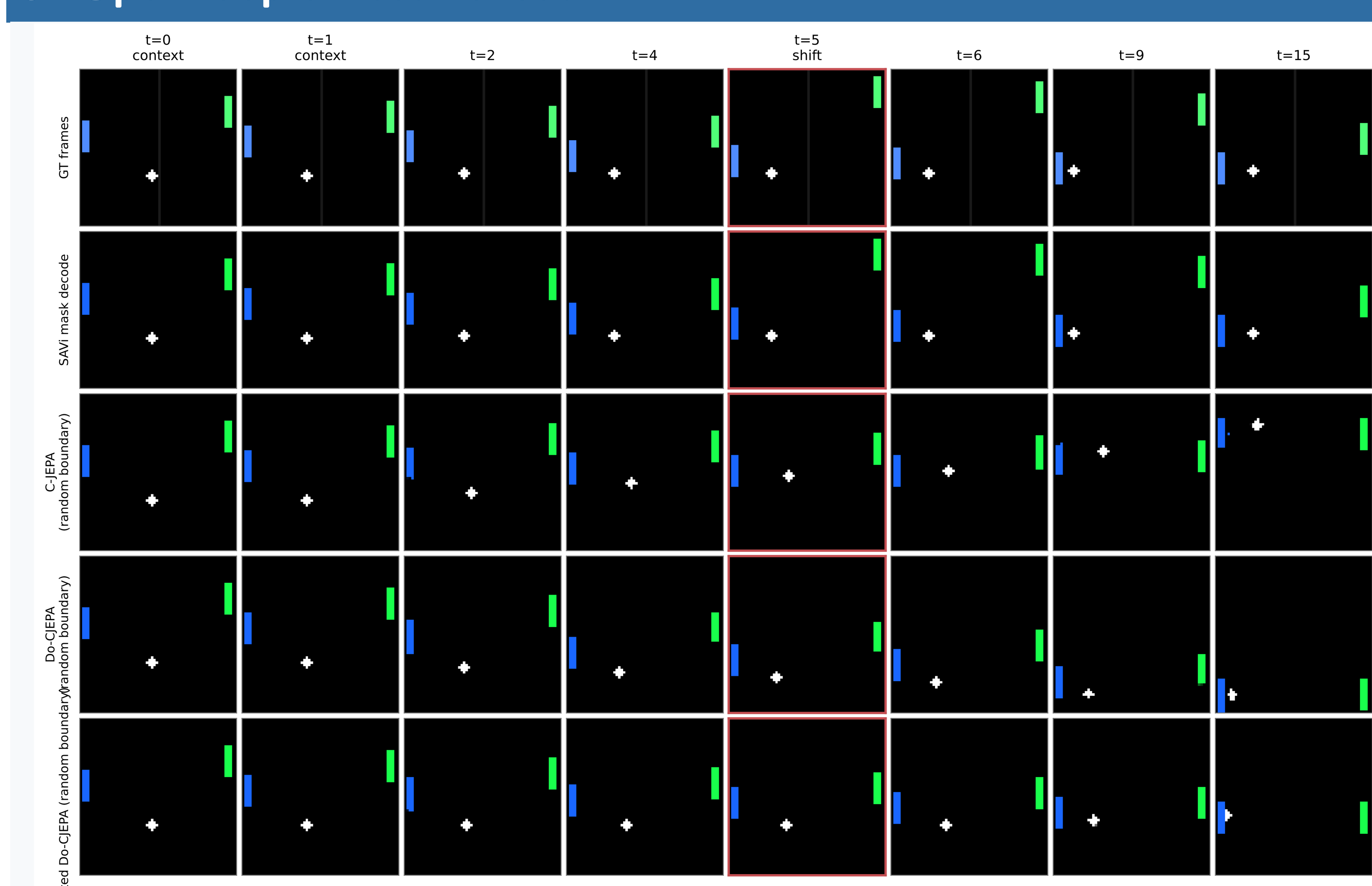


−80.1%
Do-CJEPA relative to C-JEPA

−68.5%
gated Do-CJEPA relative to C-JEPA

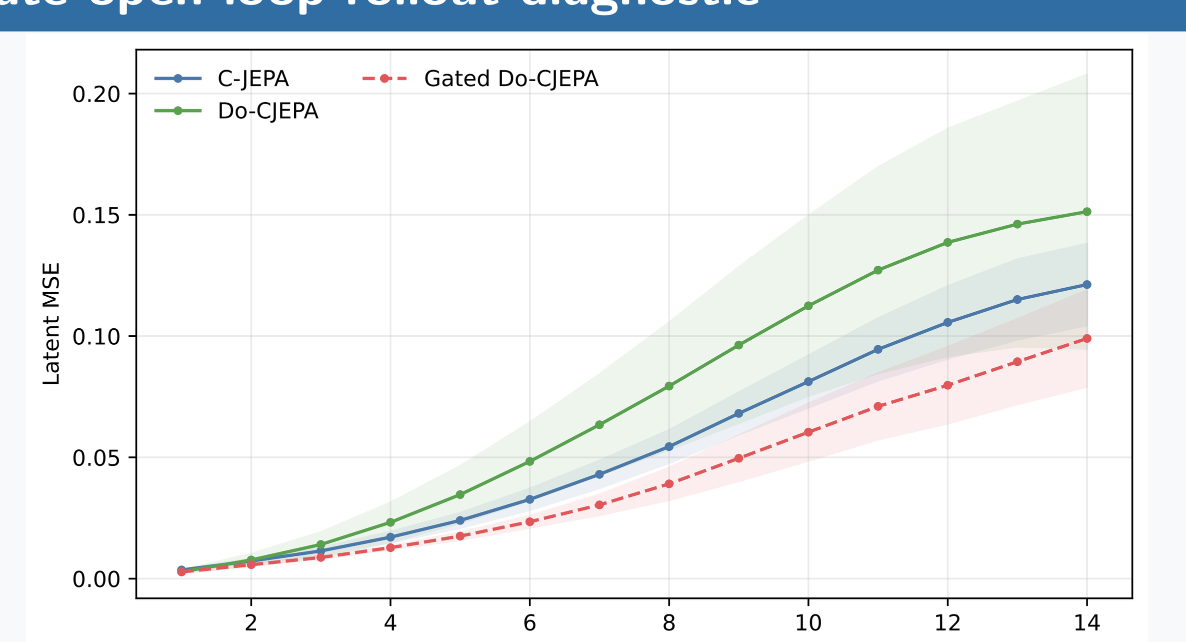
Across the first five post-shift prediction times, $t - t^* = 1, \ldots, 5$, mean invariant-token MSE is 0.01227 for C-JEPA, 0.00244 for Do-CJEPA, and 0.00387 for gated Do-CJEPA over five seeds. These values correspond to reductions of 80.1% and 68.5% relative to C-JEPA. Ordinary Do-CJEPA has the lowest error when each prediction receives the observed history through $t - 1$.

5. Open-loop shifted rollout



Top to bottom: ground truth, SAVi-mask rendering, C-JEPA, Do-CJEPA, and gated Do-CJEPA. All models receive $t = 0, 1$ and roll forward through $t = 15$. The red border marks the unknown right-paddle intervention at $t = 5$. For this sequence, MSE is 0.04956 for C-JEPA, 0.06740 for Do-CJEPA, and 0.01953 for gated Do-CJEPA. This trajectory is illustrative; Section 6 reports the aggregate comparison.

6. Aggregate open-loop rollout diagnostic



34.6%
reduction relative to Do-CJEPA

18.3%
reduction relative to C-JEPA

The token-wise gate is initialized near zero, and the response loss supervises the change in the gated causal residual. The model also uses four-step source-environment rollout training. Checkpoint selection never uses the shifted test set. At horizon 14, MSE is 0.15131 for Do-CJEPA, 0.12125 for C-JEPA, and 0.09903 for gated Do-CJEPA. Gated Do-CJEPA has lower error than C-JEPA in all five matched seeds and lower error than Do-CJEPA in four of five.

Assumptions required by the current pipeline

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

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

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

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.

References

- H. Nam, Q. Le Lidec, L. Maes, Y. LeCun, and R. Balestriero. *Causal-JEPA: Learning World Models through Object-Level Latent Masking*. 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*. NeurIPS, 37:77269–77315, 2024.