WORLDTRACES: A DATA ATTRIBUTION FRAMEWORK FOR PREDICTIVE WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Data attribution measures how much each training sample shapes what a model predicts. However, it has not been explored in predictive world models that learn representations from large-scale sequential data such as trajectories and rollouts. We remark that there is a disconnection between existing training-data attribution methods and distinct behaviors learned by a world model such as JEPA. Specifically, many scalable data attribution methods require selecting a scalar model output or optimization-based proxy, whereas JEPA's asymmetric online–target architecture does not provide a single canonical scalar quantity aligned with its predictive, representational, and temporal behaviors. To fill this gap, we propose WorldTraces to measure training data's attribution in future-latent prediction, representation consistency, and temporal latent structure. It combines these responses into a query-conditioned ranking of training examples. Although world-model architectures vary widely, we focus on JEPA because it is a leading architecture for learning general-purpose world representations and a practical test case for tracing how training data relates to learned dynamics and downstream behaviors. Experiments on V-JEPA 2 show that WorldTraces aligns more closely than adapted Grad-Sim and LESS with effects observed under subset interventions across predictive, representational, and structural behaviors. In the primary controlled IRIS setting, WorldTraces maintains a consistent ordering across these behaviors, providing evidence that it extends beyond JEPA in this setting. WorldTraces requires neither subset retraining nor stored training trajectories, and its runtime scales approximately linearly with candidate-pool size at fixed query count.

1 INTRODUCTION

World models are used in systems that must understand, predict, and eventually act in the physical world (Hafner et al., 2023). Predictive world models learn how an environment evolves from experience, including relations among states, predictable changes, and what the present implies about the future (Ha & Schmidhuber, 2018; Ding et al., 2025). These models take several forms: some predict future observations or discrete tokens, whereas Joint-Embedding Predictive Architectures (JEPAs) predict in a learned representation space. Rather than reconstruct every detail of missing or future observations, JEPAs preserve predictable semantic and temporal structure while discarding details that are difficult to predict and often irrelevant (Bardes et al., 2023; Chen et al., 2026). V-JEPA 2 scales this approach to internet-scale visual data and learns representations that support object and motion understanding, action anticipation, and future-state prediction (Assran et al., 2025). These advances raise a question for the predictive world model: which training experiences or data samples gave rise to the model's representations and predictive behavior?

This question becomes more pressing as world models and their training sets grow. Large sequential datasets may contain redundant, uninformative, biased, or corrupted samples, while data curation and downstream adaptation require developers to decide which examples to retain, inspect, or select (Paul et al., 2021; Sorscher et al., 2022). Yet a trained world model does not reveal the contribution of individual trajectories or videos to its internal representations and predictions, making it difficult to trace a behavior or failure to the training experiences that shaped it (Hara et al., 2019; Barshan et al., 2020). We call this missing connection the predictive world-model attribution gap. Conceptually, attribution asks how a measured model behavior would change if a candidate training experience were absent. Leave-one-out retraining provides a natural reference, but training for every candidate is

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

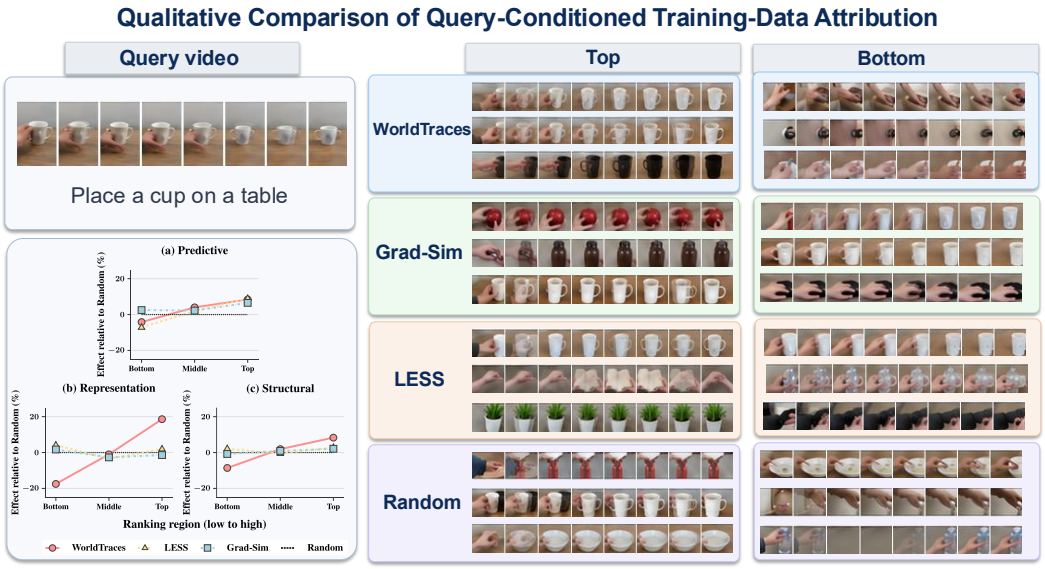

Figure 1: Illustrative comparison of query-conditioned training-data attribution. Different attribution objectives can produce substantially different training-data explanations for the same JEPA query, motivating attribution based on the learned behaviors of the world model.

infeasible (Nguyen et al., 2023; Zhang et al., 2023). The challenge is therefore to estimate candidate influence from an existing checkpoint without reconstructing every training-data counterfactual.

Prior work estimates training-data influence through approaches ranging from retraining-based marginal utility to scalable proxies derived from parameter sensitivity, optimization trajectories, gradients, or learned representations (Ghorbani & Zou, 2019; Jia et al., 2019; Ilyas et al., 2022; Koh & Liang, 2017; Schioppa et al., 2022; Pruthi et al., 2020; Park et al., 2023; Xia et al., 2024; Yeh et al., 2018). These proxies can measure optimization influence, but they do not directly capture the distinct behaviors learned by a predictive world model. The training objective determines how the model is fit, but it does not fully describe the resulting predictive, representational, and temporal behaviors. This distinction is particularly relevant to JEPA, where the masked latent-prediction loss need not reflect the quality of the learned representations (Li et al., 2026b). Different training examples may support future prediction, representation consistency, or temporal organization even when loss- or gradient-based methods assign them similar values. Figure 1 illustrates this distinction.

We treat attribution as an empirical question about learned model behavior. We remark that a useful method should predict controlled changes in the behaviors of interest, remain informative when influential data are modified, and avoid subset retraining or stored training trajectories. This motivates two questions: (1) how should training-data influence be defined for a predictive world model, and (2) can it be estimated efficiently from a trained checkpoint?

To answer the questions, we introduce WorldTraces, a checkpoint-level behavioral attribution method for predictive world models. For each candidate, WorldTraces applies a local forgetting-directed intervention to the trained checkpoint and measures responses on fixed reference queries. The intervention reverses the candidate-specific training direction as a controlled checkpoint perturbation rather than approximating exact unlearning or leave-one-out retraining. WorldTraces measures changes in future prediction, representation consistency, and temporal structure, then combines their query-conditioned ranks into a candidate ranking. The model-specific training objective determines the intervention direction, while learned behaviors, rather than training loss alone, determine influence.

Our main experiments use the action-free V-JEPA 2 model and compare WorldTraces with JEPA-adapted versions of Grad-Sim and LESS under matched checkpoints, candidates, queries, masks, and preprocessing. We evaluate grouped behavioral interventions, component and weight choices, linear datamodeling, controlled corruption, and computational cost. A controlled continued-training experiment on IRIS applies the same attribution principle with architecture-specific losses and

readouts. This experiment tests WorldTraces outside JEPA in a controlled setting, rather than claiming universal transfer across predictive world models. Our contributions are threefold:

(1) We formulate training-data attribution for predictive world models as a behavioral task: identifying examples that support prediction, representation consistency, and temporal structure.

(2) We introduce WorldTraces, which ranks candidates by their responses to targeted, forgetting-directed checkpoint interventions without candidate-wise retraining or stored training trajectories.

(3) We evaluate the method primarily on V-JEPA 2, with controlled cross-architecture validation on IRIS. Across the JEPA experiments, WorldTraces rankings track grouped-intervention effects across all three dimensions. The same ordering is observed in the primary non-JEPA evaluation setting.

2 Training-Data Attribution for JEPA World Models

2.1 V-JEPA 2 as the Attribution Target

V-JEPA 2 learns from unlabeled video by predicting masked content in a learned representation space rather than reconstructing pixels (Bardes et al., 2024; Assran et al., 2025). A video v is divided into spatiotemporal tubelets, some visible and others masked. The context encoder E_θ processes the visible tokens, the predictor P_ϕ estimates representations for masked regions, and the target encoder $E_{\bar{\theta}}$ provides the corresponding stop-gradient targets. The target encoder is updated as an exponential moving average of θ and receives no gradient. We denote the masked prediction objective for video v by $\mathcal{L}_{\mathrm{JEPA}}(v; \theta, \phi, \bar{\theta})$.

Our aim is not to attribute the JEPA training loss itself, but to identify which training examples support the pretrained model's behaviors. WorldTraces therefore evaluates attribution through changes in learned JEPA behavior, using $\mathcal{L}_{\mathrm{JEPA}}$ to construct a local candidate-specific intervention on the online encoder and predictor while keeping the target encoder fixed as a common latent reference.

2.2 Behavioral Training-Data Attribution

Let $\mathcal{D}$ denote the pretraining dataset and $\Theta_{\mathcal{D}}$ the V-JEPA 2 model trained on it. Let $\mathcal{Z} = \{z_i\}_{i=1}^{N}$ denote the candidate attribution units, each a video or fixed data shard, and let q be a query video. For $k \in \{\mathrm{P}, \mathrm{R}, \mathrm{S}\}$, $F_k(q; \Theta)$ denotes one of three behavioral readouts: future-latent prediction, representation consistency, or temporal latent structure.

In principle, candidate z_i's influence can be defined by retraining without it and measuring the resulting behavioral change (Koh & Liang, 2017; Ghorbani & Zou, 2019):

$$a_k^\star(z_i, q) = d_k\big(F_k(q; \Theta_{\mathcal{D}}), F_k(q; \Theta_{\mathcal{D}\setminus z_i})\big), \qquad k \in \{\mathrm{P}, \mathrm{R}, \mathrm{S}\}, \tag{1}$$

where d_k is a behavior-specific change measure. We measure degradation for future-latent prediction and the magnitude of change for representation consistency and temporal latent structure, for which neither direction is inherently preferable. WorldTraces responses are defined in Section 3.

Equation 1 is a conceptual leave-one-out reference, not an observable ground truth. Direct evaluation would require retraining V-JEPA 2 for every candidate, and the measured effect may vary with large-scale pretraining stochasticity (Nguyen et al., 2023). The three behavioral readouts may also rank the same candidates differently. WorldTraces does not claim to estimate this quantity directly. Instead, it uses an operational estimand defined by the behavioral response to a targeted forgetting-directed intervention at the pretrained checkpoint. This estimand requires no candidate-wise retraining and remains distinct from the causal effect of excluding a candidate from stochastic pretraining.

2.3 The Attribution Challenge

Candidate-wise V-JEPA 2 retraining is impractical at scale. Scalable attribution methods use training loss, optimization dynamics, curvature, gradients, or representation similarity as proxies (Hammoudeh & Lowd, 2024). Though useful, these proxies do not directly measure a pretrained JEPA model's predictive, representational, and temporal behaviors. V-JEPA 2's masked-prediction objective provides its training signal but does not fully characterize learned behaviors (Li et al., 2026a).

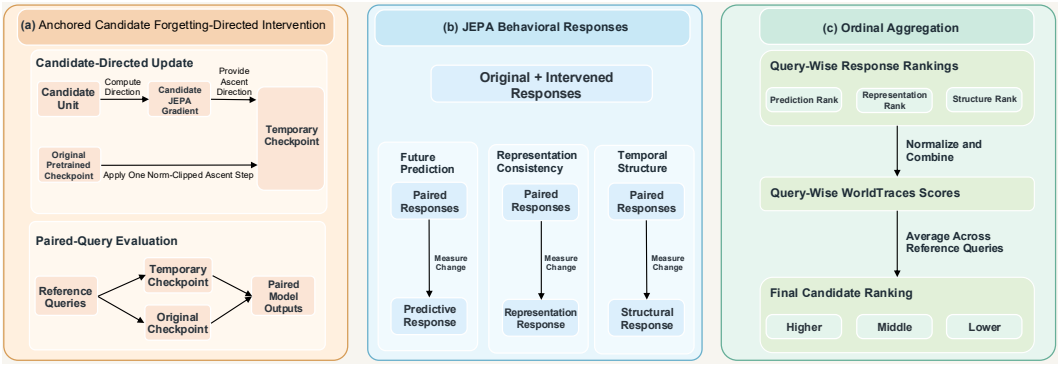

Figure 2: WorldTraces pipeline. A candidate-specific norm-clipped ascent step produces a temporary checkpoint, whose predictive, representational, and structural effects are measured against the original model on shared reference queries. Query-wise rankings are then normalized, combined, and averaged to obtain the final candidate ranking.

We seek a checkpoint-level probe linking candidate-specific, forgetting-directed interventions to measurable changes in JEPA behavior. We measure the specified intervention's effect rather than reproduce the training trajectory after candidate exclusion. We test whether the ranking predicts controlled behavioral effects, remains informative under data interventions, and requires neither candidate-wise retraining nor stored training trajectories. These tests assess effectiveness, reliability, and efficiency.

3 WorldTraces

3.1 Anchored Candidate Forgetting-Directed Intervention

Let $\omega = (\theta, \phi)$ denote the parameters of the online context encoder and predictor, and let Ω be a fixed parameter subspace. For a candidate unit z, WorldTraces computes the gradient of its average JEPA loss within Ω and takes one norm-clipped ascent step. Define $\mathrm{clip}_c(g) = g/\max(1, \|g\|_2/c)$. The intervention is

$$g_z = \Pi_\Omega \nabla_\omega \left[\frac{1}{|z|} \sum_{v \in z} \mathcal{L}_{\mathrm{JEPA}}(v; \omega, \bar{\theta}) \right], \qquad \omega_z^+ = \omega + \eta\, \mathrm{clip}_c(g_z), \qquad \Theta_z^+ = (\omega_z^+, \bar{\theta}). \quad (2)$$

Here, Π_Ω masks coordinates outside Ω, c is the clipping threshold, and η is the intervention step size. The ascent step moves opposite to a gradient-descent update on z. The target encoder remains fixed as a common latent reference. Appendix C.2 examines the sensitivity of the resulting ranking to this choice. Each perturbed checkpoint is used to evaluate all reference queries, after which the original online parameters are restored.

Because the ascent update reverses the candidate's JEPA descent direction, we treat it as a forgetting-directed checkpoint intervention rather than an arbitrary parameter perturbation. The resulting response remains a local sensitivity measure, not exact unlearning, and carries no approximation guarantee relative to leave-one-out pretraining. Paired-evaluation controls are detailed in Appendix B.

3.2 JEPA Behavioral Responses

WorldTraces measures the three behaviors introduced in Section 2.2. Predictive value records a directional change in future-latent prediction, whereas representation and structural values record unsigned changes in their respective readouts. Figure 2 gives an overview of the WorldTraces pipeline.

Predictive value. We divide the temporal tubelets of a query q into an observed prefix M_{past} and a held-out suffix M_{future}. Let $\mathcal{L}_{\mathrm{future}}(q; \omega, \bar{\theta})$ denote the V-JEPA 2 latent prediction loss on the held-out suffix, using the same latent prediction objective described in Section 2.1. We define

$$I_{\mathrm{P}}(z, q) = \left[\mathcal{L}_{\mathrm{future}}(q; \omega_z^+, \bar{\theta}) - \mathcal{L}_{\mathrm{future}}(q; \omega, \bar{\theta}) \right]_+ . \quad (3)$$

Thus, I_{P} is positive only when the intervention associated with z makes the held-out future harder to predict. The prefix-to-suffix mask is used only as an attribution-time probe and does not replace the masking objective used during V-JEPA 2 pretraining.

Representation value. Let $\mathcal{T}$ be a fixed transformation that preserves query semantics. Given N_q full-view context-encoder tokens, the pooled, unit-normalized representation is $r_\theta(q) = \mathrm{norm}\Big(N_q^{-1} \sum_{j=1}^{N_q} [E_\theta(q)]_j\Big)$. Its cross-view consistency and response to the candidate intervention are

$$V_{\mathrm{R}}(q;\theta) = r_\theta(q)^\top r_\theta(\mathcal{T}(q)), \qquad I_{\mathrm{R}}(z,q) = \left|V_{\mathrm{R}}(q;\theta) - V_{\mathrm{R}}(q;\theta_z^+)\right|. \tag{4}$$

θ_z^+ denotes the context-encoder parameters in ω_z^+. As an unsigned difference, I_{R} measures cross-view consistency sensitivity rather than directional change in representation quality. The transformation is fixed before attribution; implementation details are provided in Appendix B.

Structural value. Let $H_{t,s}^\theta(q)$ denote a full-view encoder token at temporal position t and spatial position s. With N_s spatial positions, we define the pooled latent state as $h_t^\theta(q) = N_s^{-1} \sum_{s=1}^{N_s} H_{t,s}^\theta(q)$ and its first-order temporal transition as $\delta_t^\theta(q) = h_{t+1}^\theta(q) - h_t^\theta(q)$. The structural response is

$$I_{\mathrm{S}}(z,q) = \frac{1}{(N_t-1)d} \sum_{t=1}^{N_t-1} \left\| \delta_t^{\theta_z^+}(q) - \delta_t^\theta(q) \right\|_2^2. \tag{5}$$

Here, N_t and d denote the number and dimensionality of the temporal latent states. Unlike the displacement of individual representations, I_{S} measures changes in the relations between successive latent states. It is an unsigned sensitivity measure for this first-order temporal readout, not an estimate of the environment's complete physical dynamics. Sensitivity to alternative representation transformations and temporal-structure readouts is examined in Appendix C.5.

3.3 Ordinal Aggregation

The three responses have different units and scales. For each query q and component k, let $r_{k,q}(z)$ denote the descending rank of candidate z, with average ranks assigned to ties. For $N > 1$, we normalize the rank as $\widetilde{I}_k(z,q) = 1 - \frac{r_{k,q}(z)-1}{N-1}$. This query-wise normalization makes components comparable but discards raw magnitudes. Using the main protocol's fixed equal weights, we define

$$\mathrm{WorldTraces}(z,q) = \frac{1}{3} \sum_{k \in \{\mathrm{P},\mathrm{R},\mathrm{S}\}} \widetilde{I}_k(z,q), \qquad \mathrm{WorldTraces}(z) = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathrm{WorldTraces}(z,q). \tag{6}$$

Alternative weights are considered only in the sensitivity analysis. WorldTraces is an ordinal score defined relative to its candidate pool, so its values are comparable only within a fixed candidate universe and evaluation protocol.

For N candidates and $|\mathcal{Q}|$ queries, WorldTraces requires N candidate-gradient computations and $O(N|\mathcal{Q}|)$ perturbed query evaluations, with a constant number of forward passes for each candidate-query pair. It requires no subset retraining, stored training trajectories, or Hessian inversion. Implementation details and edge-case handling are provided in the experimental protocol, while Appendix C.1 reports measured runtime scaling with both candidate-pool and query-set size, together with final-output storage.

4 Experiments

Models and data. Our primary experiments use the action-free V-JEPA 2 ViT-L/16 pretrained checkpoint (Assran et al., 2025). Candidate examples come from Mini-VM22M, which includes Something-Something-v2, Kinetics, HowTo100M, WebVid, and ImageNet. The effectiveness and reliability experiments use a fixed set of 100 reference videos drawn equally from Something-Something-v2, Kinetics, HowTo100M, and WebVid. Figure 5 analyzes ranking stability with respect to the size and source composition of this reference set. ImageNet examples remain in the candidate

K	Subset	Predictive effect ($\times 10^{-4}$) ↑	vs. Random (%)	Representation effect ($\times 10^{-6}$) ↑	vs. Random (%)	Structural effect ($\times 10^{-4}$) ↑	vs. Random (%)
1K	Top	0.085	+6.3	1.119	+16.8	2.392	+9.4
	Middle	0.081	+1.3	0.907	−5.3	2.122	−3.0
	Bottom	0.078	−2.5	0.782	−18.4	1.923	−12.1
	Random	0.080	0.0	0.958	0.0	2.187	0.0
5K	Top	0.437	+8.4	5.227	+18.6	11.478	+8.3
	Middle	0.419	+4.0	4.366	−1.0	10.799	+1.9
	Bottom	0.386	−4.2	3.639	−17.5	9.687	−8.6
	Random	0.403	0.0	4.409	0.0	10.598	0.0
10K	Top	0.844	+5.8	11.069	+24.5	23.563	+11.0
	Middle	0.803	+0.6	8.796	−1.1	21.171	−0.3
	Bottom	0.783	−1.9	7.018	−21.1	18.537	−12.7
	Random	0.798	0.0	8.894	0.0	21.234	0.0

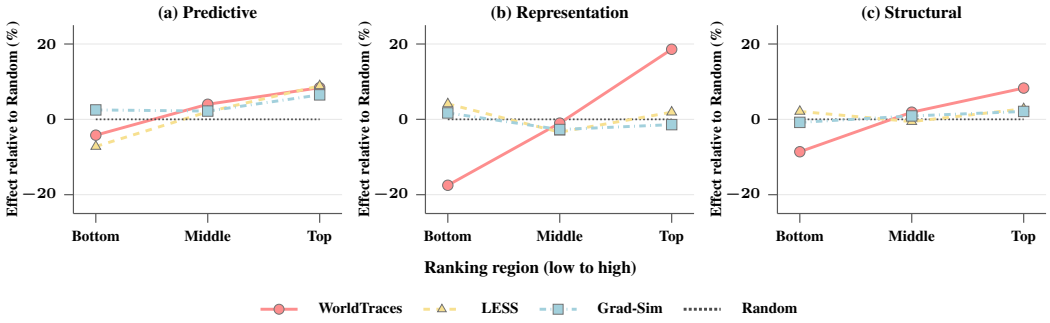

Figure 3: Upper table: grouped-intervention effects across WorldTraces ranking regions and subset sizes, with percentages relative to size-matched Random subsets. Lower panels: matched comparison with Grad-Sim and LESS at $K = 5\text{K}$, with the shared Random reference shown by the dotted line.

pool but are excluded from the query set because the predictive and structural readouts require temporal inputs. Section 4.1 evaluates cross-architecture transfer using IRIS on Breakout.

Attribution and baselines. For V-JEPA 2, WorldTraces intervenes on the final context-encoder block, its normalization layer, and the predictor, while keeping the EMA target encoder fixed; Appendix C.4 examines sensitivity to the intervention parameter subspace. Unless otherwise stated, we use the fixed 1:1:1 aggregation in Eq. 6. We compare WorldTraces with model-adapted versions of Grad-Sim and LESS (Pruthi et al., 2020; Xia et al., 2024). All matched methods within each architecture use the same checkpoint, candidates, queries, model-specific objective, and attribution parameter subspace. We discuss the additional design choices involved in adapting TRAK (Park et al., 2023) to V-JEPA 2 in Appendix A.2.

Evaluation protocol. For subset experiments, we extend Eq. 2 from an individual candidate to a subset $\mathcal{S}$. We compute the subset-specific gradient and apply one norm-clipped ascent step from the shared checkpoint. The temporary checkpoint is evaluated on common queries, then restored before the next intervention. This procedure measures the effect of a local, forgetting-directed checkpoint intervention rather than performing or reproducing leave-subset-out retraining on $\mathcal{D} \setminus \mathcal{S}$.

Additional implementation and reproducibility details for both architectures are provided in Appendix B.

4.1 Results and Analysis

We test whether candidate-level attribution scores produce a global ranking that predicts the effects of grouped checkpoint interventions, and whether it is more consistent than the adapted baselines' rankings. For each subset size K, we select size-matched subsets from the Top, Middle, and Bottom regions of the WorldTraces ranking and compare their intervention effects with a Random reference. Each subset is evaluated from the same checkpoint using the shared query set and the grouped intervention described in Section 4.

The upper table in Figure 3 shows that intervention effects follow a strict Top–Middle–Bottom ordering across all three behaviors at $K \in \{1\text{K}, 5\text{K}, 10\text{K}\}$. The Random reference generally lies

between the higher- and lower-ranked regions. Each result comes from a single grouped intervention rather than a sum of candidate-level responses. This consistent ordering shows that local candidate–query scores can be aggregated into a global ranking that is informative for subset-level interventions.

At $K = 5$K, we compare WorldTraces with JEPA-adapted Grad-Sim and LESS under the same evaluation protocol. The lower panels of Figure 3 show that WorldTraces is the only method with effects increasing from Bottom to Middle to Top across prediction, representation, and temporal structure. LESS preserves this ordering for prediction but not for representation or structure. Its predictive Top–Bottom separation is larger than that of WorldTraces, but this advantage does not extend across behaviors. Grad-Sim preserves the expected ordering only for structure, where its Bottom–Top range is substantially smaller than that of WorldTraces. A large single-readout separation therefore does not establish a consistent training-data ranking. Among evaluated methods, only WorldTraces maintains the expected ordering across all three behaviors.

Component and Weight Analysis. We examine whether the three WorldTraces components provide distinct attribution signals and how aggregation weights affect the resulting ranking. For each configuration, we recompute the global candidate ranking and evaluate its Top and Bottom subsets at $K = 5$K, with all other settings fixed. Figure 4 reports the resulting Top–Bottom separation for each behavioral readout.

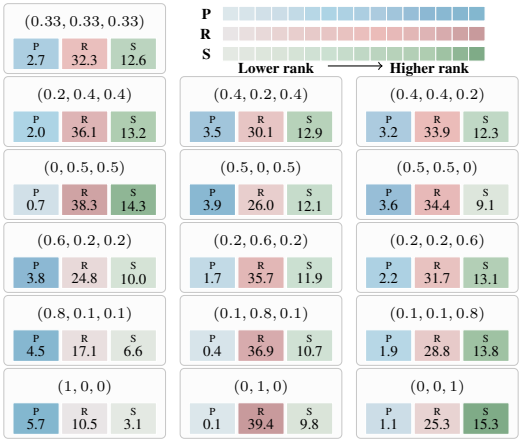

Figure 4: Top–Bottom separation (%) in predictive (P), representational (R), and structural (S) effects under different aggregation weights. Darker shading indicates a higher rank for the same behavioral readout.

Among the leave-one-component-out configurations, removing a component causes the largest reduction in its corresponding behavior. Each single-component ranking produces the strongest separation on its own readout but weaker separation on the others. The representation-only ranking retains structural separation, while the structure-only ranking retains representation separation. This suggests that the two components capture related but non-identical signals.

Along each sequence with the other two components equally weighted, increasing a component's weight consistently strengthens separation on its corresponding readout. These gains come at the expense of other behaviors, and no configuration performs best across prediction, representation, and temporal structure. We therefore retain the predefined 1:1:1 aggregation as a balanced, task-independent default rather than tuning weights to the evaluation results.

Reference-Query Sensitivity. We study how the size and source composition of the reference-query set affect the WorldTraces ranking. Using a fixed pool of 5,000 video candidates, we use the ranking from the full source-balanced set of $Q = 100$ queries as the comparison point. For each $Q \in \{4, 8, 20, 40, 60, 80\}$, we sample 50 source-balanced query sets without replacement, recompute the ranking, and measure agreement by Spearman correlation and Top-100 and Top-500 overlap.

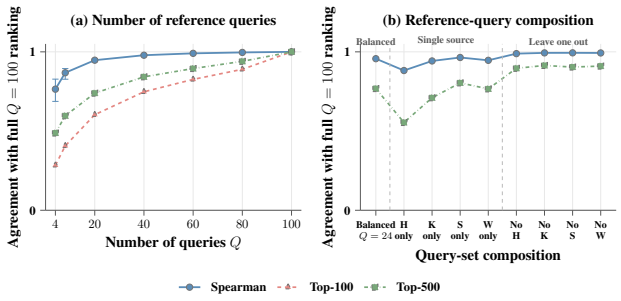

Figure 5: Reference-query sensitivity of WorldTraces. (a) Agreement with the full $Q = 100$ ranking as the number of balanced queries increases. (b) Comparison across balanced, single-source, and leave-one-source-out query sets.

Figure 5(a) shows that the global ranking becomes stable with relatively few reference queries. Mean Spearman correlation is 0.947 at $Q = 20$, 0.979 at $Q = 40$, and 0.996 at

$Q = 80$. At $Q = 80$, mean Top-500 overlap is 0.940, while Top-100 overlap is 0.890. Thus, the overall ordering is recovered reliably with a modest number of queries, although the highest-ranked candidates remain more sensitive to query sampling.

We also vary the source composition of the query set. Figure 5(b) shows that balanced $Q = 24$ subsets achieve a mean Spearman correlation of 0.957 and a mean Top-500 overlap of 0.768. Single-source correlations range from 0.882 for HowTo100M to 0.964 for Something-Something-v2, whereas leave-one-source-out rankings retain correlations between 0.989 and 0.994. Query composition therefore affects the attribution target, but the global ranking is not determined by any single source. We use all 100 source-balanced queries in the primary experiments to reduce sampling sensitivity, especially near the top of the ranking.

Linear Datamodeling. We test whether candidate-level attribution scores predict the relative effects of interventions across many subsets, beyond separating the Top and Bottom ranking regions. From a fixed pool of 50K candidates, we construct 50 source-balanced subsets of $K = 5$K and evaluate each on the same 100 reference queries. For each method, we aggregate the candidate scores within a subset and compare the result with the behavioral effect measured after one grouped gradient-ascent intervention over the entire subset. LDS is the Spearman correlation between the predicted and measured subset effects. The overall score is the arithmetic mean of the predictive, representational, and structural LDS values.

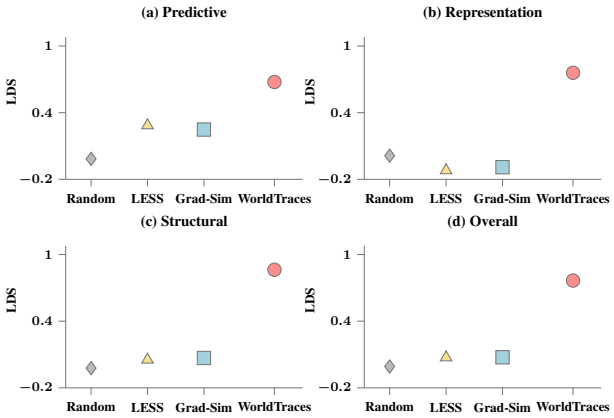

Figure 6: Linear Datamodeling Scores on 50 source-balanced subsets of $K = 5$K.

Figure 6 shows that WorldTraces achieves the highest LDS in all three behavioral dimensions, with correlations of 0.676, 0.759, and 0.863 for prediction, representation, and temporal structure, respectively. Its overall LDS is 0.766, compared with 0.075 for Grad-Sim and 0.073 for LESS. The adapted baselines have modest positive correlations for prediction, negative correlations for representation, and correlations near zero for structure. The Random reference remains near zero.

These results extend the ranking-region analysis to relative ordering across a broader collection of subsets. When aggregated within a subset, WorldTraces scores predict independently measured grouped-intervention effects. LDS therefore provides a continuous subset-level measure of attribution fidelity under the specified checkpoint intervention. It does not measure agreement with leave-subset-out retraining or with sums of individual intervention effects.

Attribution Response under Controlled Corruption. We test whether attribution scores respond more strongly to corruption of highly ranked examples than to the same corruption applied to a size-matched Random subset. Rankings are fixed before corruption. For each method, we apply the same corruption operator and

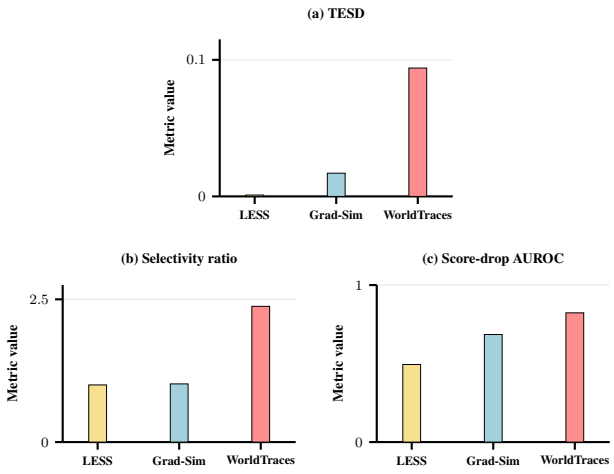

Figure 7: Controlled-corruption selectivity measured by TESD, selectivity ratio, and score-drop AUROC. No-selectivity references are 0, 1, and 0.5, respectively.

budget to its top-ranked subset and the Random reference, then recompute attribution scores using the same checkpoint and queries. We report targeted excess score drop (TESD), the ratio of targeted to random score reduction, and score-drop AUROC. Unlike the grouped-intervention evaluation above, this experiment perturbs the selected training examples rather than applying the WorldTraces checkpoint intervention as the evaluation operator.

As shown in Figure 7, WorldTraces exhibits clear selectivity across all three metrics. It achieves a TESD of 0.094, a selectivity ratio of 2.377, and a score-drop AUROC of 0.823. Its score reduction under targeted corruption is therefore more than twice that under the size-matched Random corruption. Grad-Sim has a selectivity ratio close to one but an AUROC above chance, suggesting some ability to distinguish individual examples without a larger aggregate score reduction under targeted corruption. LESS remains near the no-selectivity reference across all three metrics.

This experiment examines a different property from LDS. LDS compares attribution predictions with independently measured grouped-intervention effects, whereas controlled corruption perturbs the selected training examples directly and tests whether attribution scores respond selectively to those perturbations. It therefore provides a complementary evaluation that is decoupled from the grouped checkpoint intervention used in the LDS analysis. We do not treat this test as an independent ground truth or as a substitute for leave-subset-out retraining.

Cross-Architecture Validation on IRIS. We test whether the ranking behavior observed with V-JEPA 2 carries over to IRIS (Micheli et al., 2022), a discrete-token autoregressive world model. We use a controlled continued-training protocol so that candidate membership is known, then compare $K = 100$ subsets from each method's Top, Middle, and Bottom ranking regions with a shared Random reference. Appendix B.4 provides the training, readout, calibration, and intervention details.

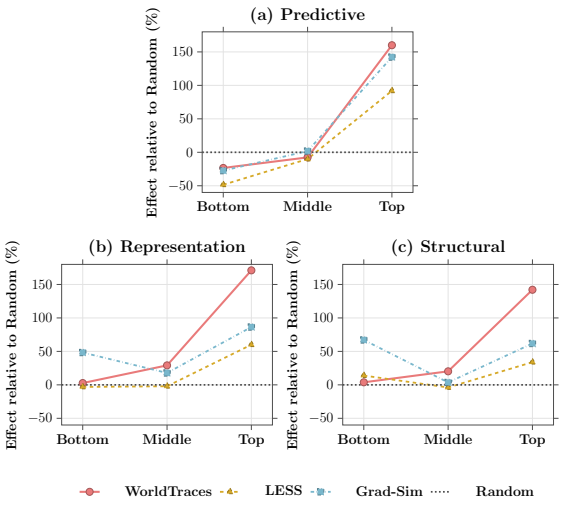

Figure 8: Controlled cross-architecture evaluation on IRIS at $K = 100$.

Figure 8 shows that the effects of WorldTraces-ranked subsets increase from Bottom to Middle to Top for prediction, representation, and temporal structure. Grad-Sim preserves this ordering only for prediction, whereas LESS preserves it for prediction and representation but not for structure. WorldTraces is the only evaluated method that maintains the expected ordering across all three behavioral readouts.

The ranking behavior observed in the primary V-JEPA 2 experiments therefore also appears in a discrete-token autoregressive world model. This result provides controlled evidence that the attribution principle is not specific to JEPA. Its scope remains limited to the evaluated IRIS checkpoint, environment, and continued-training protocol; robustness analyses are provided in Appendix C.3.

5 Conclusion

We introduced WorldTraces, a checkpoint-level behavioral attribution method that links the behavior of predictive world models to their training examples. WorldTraces applies local, forgetting-directed interventions and measures their effects on future prediction, representation consistency, and temporal structure. On V-JEPA 2, WorldTraces rankings consistently predict grouped-intervention effects across subset sizes, match measured subset effects more closely than adapted Grad-Sim and LESS, and respond more selectively to targeted corruption. In the primary controlled study on IRIS, WorldTraces retains the same cross-behavior ordering in a discrete-token autoregressive world model, providing evidence that the attribution principle extends beyond JEPA in the evaluated setting. WorldTraces requires neither candidate-wise retraining nor stored optimization histories, making controlled checkpoint interventions a practical way to study training-data influence.

AI USE STATEMENT

In this work, we used generative AI tools to summarize and compare related literature, assist with drafting and revising parts of the manuscript, improve readability and organization, and suggest presentation choices for figures. All AI-assisted material was reviewed and edited by the authors. The authors take responsibility for the final content of this work.

ETHICS STATEMENT

This work does not involve human-subject studies or the collection of personally identifiable information. Our experiments use existing research datasets, pretrained model checkpoints, and simulated environments to study training-data attribution in predictive world models. WorldTraces is intended as an analysis tool for understanding how training examples influence learned model behavior. As with other data-attribution methods, its outputs may reflect biases or limitations present in the underlying data and models. Attribution scores should therefore be interpreted with care when used for data curation, removal, or model auditing. We are not aware of additional ethical concerns specific to the methodology beyond those associated with the underlying datasets and models.

REPRODUCIBILITY STATEMENT

The main text provides the full definition of WorldTraces, including the candidate-specific intervention, behavioral responses, and score aggregation in Sections 2–4. Appendix B contains the implementation and experimental details needed to reproduce the reported results, including data preprocessing, the intervention procedure, baseline implementations, the IRIS protocol, matched evaluation settings, random seeds, and computing environment. Section 4 reports the primary evaluation and reference-query sensitivity analysis, while Appendix C provides additional analyses of computational efficiency, target-encoder coupling, IRIS robustness, intervention parameter-subspace sensitivity, and readout-definition sensitivity. Candidate and query manifests are fixed before evaluation and reused across matched methods, and the reported efficiency measurements are collected under a common hardware and software setting. Together, these sections specify the experimental protocol and evaluation conditions used throughout the paper.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629. IEEE, 2023.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Adrien Bardes, Jean Ponce, and Yann LeCun. Mc-jepa: A joint-embedding predictive architecture for self-supervised learning of motion and content features. *arXiv preprint arXiv:2307.12698*, 2023.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Elnaz Barshan, Marc-Etienne Brunet, and Gintare Karolina Dziugaite. Relatif: Identifying explanatory training samples via relative influence. In *International Conference on Artificial Intelligence and Statistics*, pp. 1899–1909. PMLR, 2020.

Samyadeep Basu, Xuchen You, and Soheil Feizi. On second-order group influence functions for black-box predictions. In *International Conference on Machine Learning*, pp. 715–724. PMLR, 2020.

Delong Chen, Mustafa Shukor, Theo Moutakanni, Willy Chung, Lei Yu, Tejaswi Kasarla, Allen Bolourchi, Yann LeCun, and Pascale Fung. Vl-jepa: Joint embedding predictive architecture for vision-language. In *International Conference on Learning Representations*, volume 2026, pp. 89791–89807, 2026.

Zheng Dai and David K Gifford. Training data attribution for diffusion models. *arXiv preprint arXiv:2306.02174*, 2023.

Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, et al. Understanding world or predicting future? a comprehensive survey of world models. *ACM Computing Surveys*, 58(3):1–38, 2025.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, pp. 2242–2251. PMLR, 2019.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3):440, 2018.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Zayd Hammoudeh and Daniel Lowd. Training data influence analysis and estimation: A survey. *Machine Learning*, 113(5):2351–2403, 2024.

Satoshi Hara, Atsushi Nitanda, and Takanori Maehara. Data cleansing for models trained with sgd. *Advances in neural information processing systems*, 32, 2019.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Predicting predictions from training data. *arXiv preprint arXiv:2202.00622*, 2022.

Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J Spanos. Towards efficient data valuation based on the shapley value. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1167–1176. PMLR, 2019.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. Pmlr, 2017.

Yongchan Kwon and James Zou. Beta shapley: a unified and noise-reduced data valuation framework for machine learning. *arXiv preprint arXiv:2110.14049*, 2021.

Yongchan Kwon, Eric Wu, Kevin Wu, and James Y Zou. Datainf: Efficiently estimating data influence in lora-tuned llms and diffusion models. In *International Conference on Learning Representations*, volume 2024, pp. 21921–21942, 2024.

Yann LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Xianhang Li, Chen Huang, Chun-Liang Li, Eran Malach, Josh Susskind, Vimal Thilak, and Etai Littwin. Rethinking JEPA: Compute-efficient video SSL with frozen teachers. In *International Conference on Learning Representations*, 2026a.

Xianhang Li, Chen Huang, Chun-Liang Li, Eran Malach, Joshua Susskind, Vimal Thilak, and Etai Littwin. Rethinking jepa: Compute-efficient video self-supervised learning with frozen teachers. In *International Conference on Learning Representations*, volume 2026, pp. 142610–142633, 2026b.

Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. *arXiv preprint arXiv:2209.00588*, 2022.

Elisa Nguyen, Minjoon Seo, and Seong Joon Oh. A bayesian approach to analysing training data attribution in deep learning. *Advances in Neural Information Processing Systems*, 36:64155–64180, 2023.

Nicolas Papernot and Patrick McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *arXiv preprint arXiv:1803.04765*, 2018.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. Trak: Attributing model behavior at scale. *arXiv preprint arXiv:2303.14186*, 2023.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in neural information processing systems*, 34: 20596–20607, 2021.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. *Advances in neural information processing systems*, 33: 19920–19930, 2020.

Andrea Schioppa, Polina Zablotskaia, David Vilar, and Artem Sokolov. Scaling up influence functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8179–8186, 2022.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536, 2022.

Sheng-Yu Wang, Aaron Hertzmann, Alexei A Efros, Jun-Yan Zhu, and Richard Zhang. Data attribution for text-to-image models by unlearning synthesized images. *Advances in Neural Information Processing Systems*, 37:4235–4266, 2024.

Xindi Wu, Despoina Paschalidou, Jun Gao, Antonio Torralba, Laura Leal-Taixé, Olga Russakovsky, Sanja Fidler, and Jonathan Lorraine. Motion attribution for video generation. *arXiv preprint arXiv:2601.08828*, 2026.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. Less: Selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333*, 2024.

Tong Xie, Haoyu Li, Andrew Bai, and Cho-Jui Hsieh. Data attribution for diffusion models: Timestep-induced bias in influence estimation. *arXiv preprint arXiv:2401.09031*, 2024.

Chih-Kuan Yeh, Joon Kim, Ian En-Hsu Yen, and Pradeep K Ravikumar. Representer point selection for explaining deep neural networks. *Advances in neural information processing systems*, 31, 2018.

Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. Counterfactual memorization in neural language models. *Advances in Neural Information Processing Systems*, 36:39321–39362, 2023.

A Related Work

A.1 Joint-Embedding Predictive World Models

Joint-Embedding Predictive Architectures (JEPAs) predict target representations from partial context rather than reconstructing observations in pixel space (LeCun et al., 2022; Assran et al., 2023; Bardes et al., 2024). I-JEPA established this approach for images using a context encoder, a predictor, and a momentum-updated target encoder (Assran et al., 2023). V-JEPA extended it to video and showed that latent-space prediction can produce representations sensitive to motion and appearance without negative pairs, text supervision, or pixel reconstruction (Bardes et al., 2024). V-JEPA 2 subsequently scaled action-free JEPA pretraining to larger and more diverse video datasets, supporting understanding, anticipation, and prediction. An action-conditioned variant was then post-trained for planning (Assran et al., 2025).

The architecture also shapes how training-data value should be defined. A V-JEPA 2 example may contribute to predicting unobserved latent states, maintaining representation consistency, or organizing relations between successive latent states. These effects need not align with its contribution to the scalar pretraining loss, which may not reliably indicate the quality of the learned representations (Li et al., 2026a). We therefore treat the complete action-free V-JEPA 2 system, including the context encoder, predictor, and EMA target encoder, as the attribution target. WorldTraces does not introduce a new JEPA training objective or attribute the downstream action-conditioned controller. It is a post-hoc sensitivity method for studying how pretraining candidates support complementary behaviors in a pretrained JEPA model.

A.2 Training-Data Attribution at Scale

Classical data valuation measures contribution through changes in utility when training examples are added to or removed from a dataset. Shapley-based formulations provide a principled marginal-utility interpretation, but require training on many data subsets and have therefore motivated less expensive approximations (Ghorbani & Zou, 2019; Jia et al., 2019; Kwon & Zou, 2021). Datamodels take a related subset-based approach by learning a linear relationship between training-set membership and model outputs (Ilyas et al., 2022). These methods provide natural counterfactual references, but repeated subset training is difficult at the scale of V-JEPA 2 and generally requires a predefined output or task-level utility.

Influence functions approximate leave-one-out effects using local curvature around the learned parameters (Koh & Liang, 2017). Later work extends this formulation to groups of examples and improves its scalability (Basu et al., 2020; Schioppa et al., 2022). DataInf provides an efficient closed-form influence approximation for parameter-efficient fine-tuning (Kwon et al., 2024). Although these methods avoid explicit leave-one-out retraining, they depend on local approximations to non-convex training and usually measure influence through a selected scalar objective.

Another line of scalable methods traces relationships between training and query gradients during optimization. Early work estimates influence by reconstructing SGD updates (Hara et al., 2019), while TracIn accumulates gradient alignment over saved checkpoints (Pruthi et al., 2020). TRAK uses random projections and a kernel-based formulation to approximate counterfactual model behavior at scale (Park et al., 2023). LESS combines optimizer-aware gradient features with random projection to retrieve influential instruction-tuning examples efficiently (Xia et al., 2024). These methods provide practical attribution signals, but alignment under a selected training loss need not reflect every behavior encoded by a pretrained JEPA system.

Representation-based approaches identify explanatory training examples through learned-feature contributions or proximity (Yeh et al., 2018; Papernot & McDaniel, 2018). They can retrieve semantically related examples without reproducing the training process, but similarity or representational contribution alone does not establish how perturbing a candidate would affect model behavior. Attribution estimators also differ in their definitions, assumptions, and sensitivity to training randomness (Hammoudeh & Lowd, 2024; Nguyen et al., 2023). These methods provide the foundation for our study, but none directly defines attribution through the multiple latent behaviors of a pretrained JEPA model.

For a controlled comparison, our Grad-Sim and LESS baselines use the same V-JEPA 2 checkpoint, candidates, queries, masks, and preprocessing as WorldTraces. The main distinction is the attribution target. The baselines measure relationships between candidates and queries through the JEPA training loss, whereas WorldTraces constructs a candidate-specific local intervention and measures predictive, representational, and temporal-structural responses. Our linear datamodeling experiment separately tests whether each ranking predicts measured subset interventions, without treating agreement with WorldTraces as ground truth.

Considerations for applying TRAK to V-JEPA 2. TRAK can be applied in principle to differentiable predictive models, but adapting it to V-JEPA 2 requires several choices that are not specified in the original formulation. TRAK constructs projected features from gradients of a chosen scalar model output and, in its standard ensemble formulation, combines information across multiple trained model states. The original work also shows that checkpoints from a single training trajectory can be used instead of independently trained models (Park et al., 2023). In V-JEPA 2, the predictor produces a set of latent target representations rather than a canonical scalar prediction, the targets are generated by a stop-gradient EMA encoder, and the pretraining loss depends on randomly sampled masks. Applying TRAK to JEPA therefore requires choices about how to reduce the latent prediction to a scalar, how to handle the EMA target pathway across model states, and how to account for stochastic masks when constructing projected gradients. We therefore do not include TRAK as a matched checkpoint-level baseline in the primary V-JEPA 2 comparison.

A.3 Attribution for Generative and Video Models

Training-data attribution has also been adapted to generative models, whose outputs and training objectives require specialized methods. Encoded ensembles estimate how removing an example

changes the output of a diffusion model by training models on structured data subsets (Dai & Gifford, 2023). Diffusion-TracIn and Diffusion-ReTrac identify timestep-induced bias in diffusion gradient attribution and introduce a normalized score for more query-specific retrieval (Xie et al., 2024). Unlearning-based attribution instead suppresses a synthesized output and identifies training images whose losses change in response (Wang et al., 2024). These studies show that a generic attribution score may overlook properties specific to an architecture, training objective, or output space.

Motive is most closely related to our study in its focus on video-specific behavior. It uses motion-weighted loss masks to separate temporal dynamics from static appearance in gradient-based attribution for video generation (Wu et al., 2026). WorldTraces shares the view that whole-video or scalar-loss attribution can conceal behavior-specific effects, but considers a different model class and counterfactual. Motive attributes generated motion in a diffusion model using spatially weighted training gradients. WorldTraces instead probes a non-generative V-JEPA 2 model with asymmetric online and target pathways, measuring changes in future-latent prediction, transformation consistency, and transitions between successive latent states after a local candidate-specific intervention. Motive identifies clips associated with generated motion, whereas WorldTraces studies how pretraining data support several internal behaviors of a JEPA world model.

B Additional Experimental Details

This section provides implementation and reproducibility details omitted from Section 4. Unless noted otherwise, Sections B.1–B.3 describe the primary V-JEPA 2 experiments, while Section B.4 describes the separate IRIS protocol. Sections B.5 and B.6 apply across architectures where relevant.

B.1 Data and Preprocessing

We use the action-free V-JEPA 2 ViT-L/16 checkpoint rather than the action-conditioned V-JEPA 2-AC model. Candidate and query manifests are fixed before each experiment and reused across matched methods and runs.

We follow the preprocessing pipeline of the pretrained V-JEPA 2 checkpoint for video decoding, temporal sampling, spatial preprocessing, and normalization. In paired evaluations, the original and perturbed checkpoints receive the same query clips, masks, and deterministic transformations, so the measured differences do not reflect changes in query sampling or preprocessing.

B.2 WorldTraces Intervention

The intervention is applied to the final context-encoder block, its normalization layer, and the predictor. The EMA target encoder remains fixed, providing the same latent reference for the original and perturbed checkpoints.

Each candidate intervention starts from the same pretrained checkpoint. WorldTraces computes the candidate gradient, applies the norm-clipped ascent step in Eq. 2, and evaluates all reference queries using the temporary checkpoint. The original online parameters are restored before processing the next candidate, so interventions do not accumulate across candidates.

The predictive response uses the V-JEPA 2 latent-prediction objective with an observed-prefix and held-out-suffix mask used only during attribution. For the representation response, the canonical transformation $\mathcal{T}$ is a deterministic center crop retaining 87.5% of the spatial extent, followed by bilinear resizing to the original resolution. The same transformed query is used for the original and perturbed evaluations. The structural response is computed from the corresponding full-view encoder states using the first-order temporal differences defined in Section 3.2.

B.3 Baseline Implementations

Grad-Sim computes the cosine similarity between candidate and query gradients of the V-JEPA 2 latent-prediction objective. For LESS, we retain the optimizer-aware gradient representation and similarity rule of the original method but replace its language-model objective with the V-JEPA 2 objective. These adaptations change only the model-specific objective and leave each baseline's scoring rule unchanged.

The Random reference is sampled with seed 42. Random subsets are source-balanced in experiments that impose per-source quotas.

B.4 IRIS Cross-Architecture Protocol

Controlled continued training. We use the official IRIS checkpoint for BreakoutNoFrameskip-v4 as the initialization. With seed 42, we collect $1{,}000$ candidate segments and 100 query segments, each spanning 20 consecutive timesteps. The candidate and query sets are sampled from disjoint episodes. Starting from the official checkpoint, we train the world model for one epoch on the complete candidate set using AdamW with a batch size of 64, a learning rate of 10^{-4}, weight decay of 0.01, and a maximum training-gradient norm of 10. Only the world model is updated; the visual tokenizer and actor–critic remain frozen. The query set is not used during continued training.

This construction ensures that every candidate is a known training example of the continued checkpoint, while the queries remain held out from optimization. The experiment therefore evaluates training-data attribution for the controlled continued-training stage. It does not assume that the candidates belong to the unknown training set of the original IRIS checkpoint.

Attribution subspace. All methods operate within the same fixed parameter subspace, consisting of the final Transformer block, final layer normalization, and the observation, reward, and termination prediction heads. This choice restricts attribution and intervention to late model components used by the three behavioral readouts. The subspace contains 30 parameter tensors and approximately 1.12 million parameters, corresponding to 13.26% of the world model. Parameters outside this subspace remain unchanged. WorldTraces, Grad-Sim, LESS, and the grouped evaluation all use this same parameter subspace.

Behavioral readouts. The predictive readout conditions on the first 10 timesteps of a query and measures the increase in negative log-likelihood over the remaining suffix. Negative changes are clipped to zero. The representation readout measures the change in hidden-state consistency between each query and a fixed brightness-transformed version obtained by scaling pixel values by 0.9. The structural readout measures the squared change in transitions between consecutive pooled hidden states. The world model remains in FP32, while the representation and structural reductions are accumulated in FP64 to avoid numerical cancellation at the small response scales produced by local interventions.

Intervention calibration. We select the intervention scale before running the full attribution experiment. A separate calibration uses eight candidates and 16 queries to compare step sizes of 10^{-3} and 3×10^{-3}. We examine response distinguishability, signed predictive changes, relative parameter updates, and cross-step ranking stability. We select 3×10^{-3} because it provides more stable behavioral responses while keeping the parameter update local. Global-norm clipping is fixed at 10. These settings are selected before the full $N = 1{,}000$, $Q = 100$ experiment. The same grouped-intervention settings are used to evaluate subsets selected by all methods.

WorldTraces and adapted baselines. For each candidate, WorldTraces applies one candidate-specific forgetting-directed intervention and evaluates the predictive, representational, and structural responses on the shared query set. The original parameters are restored before the next candidate is processed. Within each query, candidate responses are converted to descending, tie-aware normalized ranks. The three components are combined using the equal-weight arithmetic mean in Eq. 6, and the resulting scores are averaged across queries to produce a global candidate ranking.

Grad-Sim computes cosine similarity between candidate and query gradients of the IRIS world-model loss. LESS uses the same gradients after projection into a fixed 256-dimensional Gaussian subspace. Both baselines use the same continued checkpoint, query set, loss, and attribution parameter subspace as WorldTraces.

Grouped evaluation. For each method, the Top, Middle, and Bottom subsets contain $K = 100$ candidates selected from its global ranking. A shared Random subset of the same size provides the reference intervention. For each selected subset, we compute candidate-specific gradients using deterministic sample seeds, average the gradients, and apply one grouped forgetting-directed update

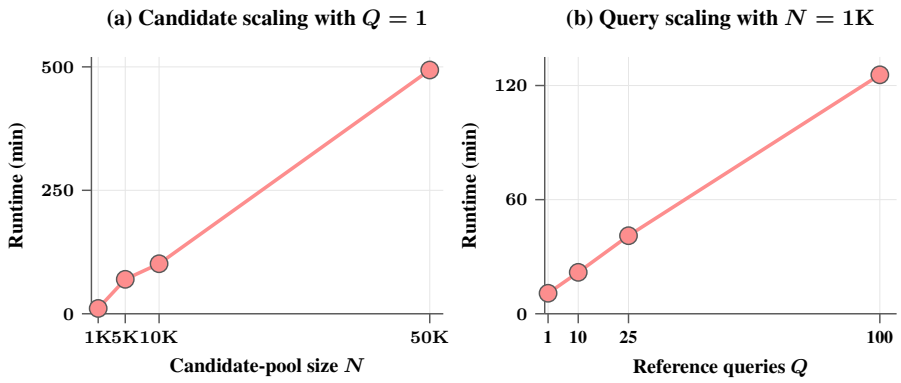

Figure 9: Runtime scaling of WorldTraces. Panel (a) varies the candidate-pool size from 1K to 50K at fixed $Q = 1$, while panel (b) varies the number of reference queries from 1 to 100 at fixed $N = 1$K.

with the calibrated step size and clipping threshold. All methods use the same grouped-update rule and hyperparameters, although the realized update norm may vary with the selected subset.

Every grouped intervention starts from the same continued checkpoint. After evaluating all 100 queries, the affected parameters are restored and checked before the next subset is processed. The reported effects are measured on the temporarily intervened model rather than obtained by summing candidate-level responses. This procedure measures a local checkpoint-level intervention effect and does not reproduce leave-subset-out retraining.

B.5 Matched Evaluation

Within each architecture, we fix the candidate and query manifests before evaluation and reuse them for all attribution methods. The methods also share the same checkpoint, query set, intervention parameter subspace, and behavioral readouts. We select subsets from the Top, Middle, Bottom, or Random regions according to the size and source constraints of each experiment. Each subset is evaluated with a single grouped forgetting-directed checkpoint intervention rather than by summing candidate-level effects. The main text reports the candidate-pool size, intervention subset size, corruption budget, and evaluation metrics for each experiment.

B.6 Randomness and Computing Environment

We use fixed seeds for randomized parts of the evaluation. The shared Random baseline uses seed 42, and frozen manifests are reused across matched methods. Efficiency measurements are collected in the same software environment and on the same hardware. All reported efficiency experiments use a single NVIDIA A100-SXM4 GPU with 80 GB of memory.

Storage measurements distinguish method-specific persistent artifacts from final attribution outputs; both exclude the shared pretrained checkpoint and raw datasets. Runtime is measured end to end under the candidate and query settings reported for each experiment.

C Additional Experimental Results

C.1 Efficiency

We measure WorldTraces at nine settings spanning $N \in \{1\text{K}, 5\text{K}, 10\text{K}, 50\text{K}\}$ and $Q \in \{1, 10, 25, 100\}$, including the joint-scale settings $(10\text{K}, 25)$ and $(50\text{K}, 10)$. All runs use the same single-A100 hardware, software environment, and timing boundary.

At fixed $Q = 1$, runtime is 10.78, 69.77, 101.30, and 493.62 minutes for $N = 1$K, 5K, 10K, and 50K, respectively. Runtime is approximately linear in N over this range ($R^2 = 0.998$). At fixed

$N = 1$K, runtime increases from 10.78 minutes at $Q = 1$ to 125.63 minutes at $Q = 100$. The two joint-scale settings take 396.95 minutes at $(10\text{K}, 25)$ and 1078.87 minutes at $(50\text{K}, 10)$. An empirical model with an NQ interaction predicts each held-out joint-scale point within approximately 4%, consistent with the N candidate-gradient computations and $O(N|Q|)$ perturbed query evaluations in Section 3.3.

Table 1: Final-output storage for the nine measured WorldTraces settings.

N	Q	Output (MiB)
1K	1	0.26
5K	1	1.32
10K	1	2.65
50K	1	13.32
1K	10	2.74
1K	25	6.86
1K	100	27.52
10K	25	69.25
50K	10	139.34

Table 1 reports final-output storage, which ranges from 0.26 MiB to 139.34 MiB across the measured settings. We do not measure $N = 500$K directly. Extrapolating the validated runtime model gives approximately 82 GPU-hours for $(500\text{K}, 1)$ and 179 GPU-hours for $(500\text{K}, 10)$ on a single A100. These values are projections rather than measured runtimes.

C.2 Target-Encoder Coupling Sensitivity

WorldTraces keeps the EMA target encoder fixed during each forgetting-directed intervention, allowing the original and perturbed online models to be evaluated against the same latent reference. We examine how sensitive the attribution ranking is to this choice. For corresponding online and target-encoder parameters, let $\Delta\theta_z = \theta_z^+ - \theta$ denote the encoder displacement induced by candidate z. We define the coupled target as $\bar{\theta}_{z,\alpha}^+ = \bar{\theta} + \alpha\Delta\theta_z$. We compare the default fixed-target setting $\alpha = 0$ with an EMA-consistent coupling $\alpha = 1 - m = 0.00075$, where $m = 0.99925$ is the target-encoder momentum, and a fully coupled setting $\alpha = 1$. The EMA-consistent setting applies this coupling only to the local intervention displacement rather than simulating an EMA training trajectory. The fully coupled setting serves as an extreme sensitivity test.

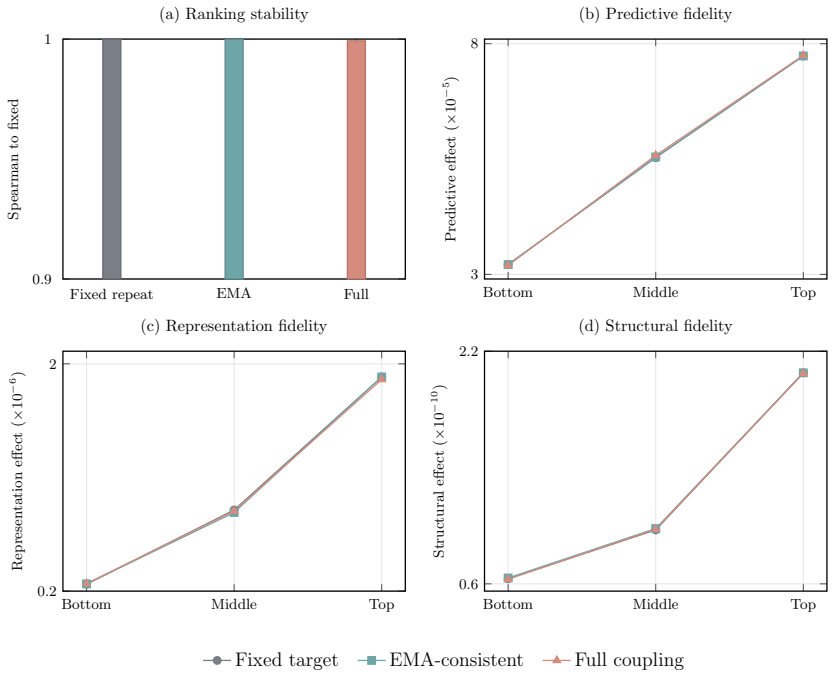

Figure 10: Target-encoder coupling sensitivity. Panel (a) reports ranking agreement with the fixed-target reference. Panels (b)–(d) compare grouped-intervention fidelity across the three behavioral readouts. All subsets are evaluated with the same fixed-target grouped-intervention protocol.

Figure 10(a) shows that the ranking changes little under target-encoder coupling. An independent fixed-target rerun has a Spearman correlation of 0.999879 with the reference ranking, compared with 0.999877 under EMA-consistent coupling. Both have a median rank displacement of one position. Full coupling produces a larger shift, but the overall ordering remains stable, with a Spearman correlation of 0.999263 and a median rank displacement of three positions.

We further examine whether these ranking changes affect grouped-intervention fidelity. For each coupling condition, we select $K = 100$ candidates from the Bottom, Middle, and Top regions of its global ranking and evaluate the resulting subsets on the same 100 reference queries using the shared fixed-target grouped-intervention protocol. Figure 10(b)–(d) shows the same Bottom–Middle–Top ordering under all three coupling conditions for the predictive, representation, and structural responses.

The differences between coupling conditions are small. Relative to the fixed-target ranking, EMA-consistent coupling changes the predictive, representation, and structural Top–Bottom separations by $+0.10\%$, -0.18%, and -0.36%, respectively. Full coupling changes them by $+0.77\%$, -1.53%, and -0.05%. These changes are small and do not follow a consistent direction across the three behavioral responses.

Target-encoder coupling therefore has little effect on either the global ranking or its grouped-intervention ordering. We retain the fixed-target formulation because it provides a common latent reference, while allowing the target encoder to move does not produce a consistent improvement in grouped-intervention fidelity.

C.3 IRIS Robustness to Forecasting Objective and Intervention Calibration

The primary IRIS experiment uses a joint objective for predicting observations, rewards, and termination signals, with intervention step size $\eta = 3 \times 10^{-3}$ and global-norm clipping threshold $c = 10$. We test whether the cross-architecture result is sensitive to either the forecasting objective or the intervention calibration.

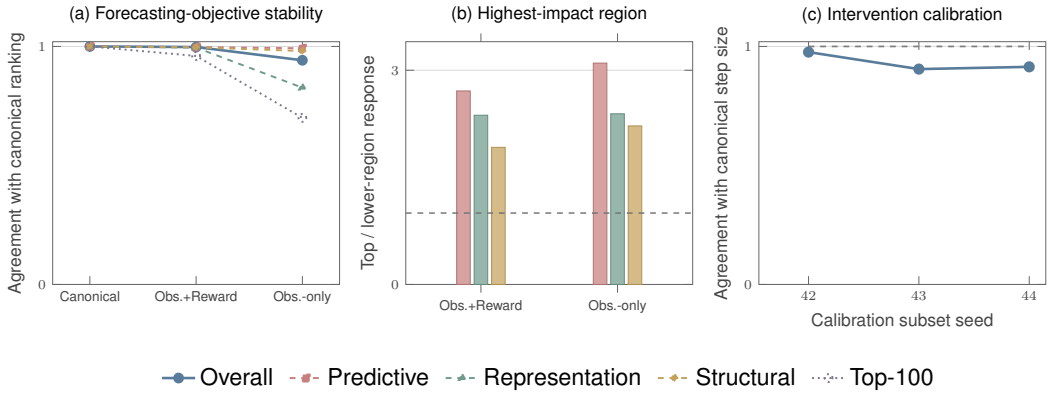

Figure 11: Robustness analysis for the IRIS evaluation. (a) Agreement with the canonical observation–reward–termination ranking under restricted forecasting objectives. (b) Separation of the highest-impact region under the two restricted objectives. Each bar gives the response of the Top $K = 100$ subset divided by the larger response of the corresponding Middle and Bottom subsets; values above one indicate a larger effect for the Top region. (c) Agreement with the canonical intervention step size $\eta = 3 \times 10^{-3}$ when using $\eta = 10^{-3}$ on three independently sampled calibration subsets. Clipping thresholds $c \in \{5, 10, 20\}$ give identical scores because clipping is inactive over the tested range.

Forecasting-objective sensitivity. IRIS predicts observations, rewards, and termination signals. To test the dependence of the attribution ranking on this objective, we recompute WorldTraces with two restricted variants: observation–reward prediction and observation-only prediction. The attribution subspace is adjusted to match each objective, while the continued checkpoint, candidate and query pools, behavioral readouts, and aggregation rule are held fixed.

Figure 11(a) shows that removing termination prediction changes the ranking very little. The observation–reward variant has Spearman correlation $\rho = 0.9975$ with the canonical ranking and retains 96% of its Top-100 candidates. The component-level correlations are similarly high: $\rho_P = 0.9975$, $\rho_R = 0.9952$, and $\rho_S = 0.9961$. Using observation prediction alone produces a larger shift, with global correlation $\rho = 0.9419$ and Top-100 overlap of 70%. Most of this change appears in the representation ranking ($\rho_R = 0.8267$); the predictive and structural rankings remain more stable, with $\rho_P = 0.9923$ and $\rho_S = 0.9809$.

We also ask whether these ranking changes alter the behavioral result of the IRIS experiment. For each restricted objective, we reselect the Top, Middle, and Bottom $K = 100$ subsets and evaluate them with the same canonical grouped-intervention operator. Figure 11(b) shows that the Top region has the largest response for all three behavioral readouts. With the observation–reward objective, the predictive, representational, and structural responses of the Top subset are $2.71\times$, $2.37\times$, and $1.92\times$ the larger response of the two lower-ranked regions. With the observation-only objective, the corresponding ratios are $3.10\times$, $2.39\times$, and $2.22\times$. The relative ordering of the Middle and Bottom regions is not preserved for every readout, indicating that the forecasting objective can affect the finer structure of the ranking. The highest-impact region, however, remains consistent across the tested objectives.

Intervention-calibration robustness. We evaluate the intervention step size and clipping threshold on three independently sampled calibration subsets, each containing eight candidates and 16 held-out queries. Within each subset, we compare $\eta = 10^{-3}$ with the primary setting $\eta = 3 \times 10^{-3}$ and vary the global-norm clipping threshold over $c \in \{5, 10, 20\}$. Ranking comparisons are always made within the same candidate and query subset.

As shown in Figure 11(c), the rankings remain similar across the three calibration subsets when the step size is reduced from 3×10^{-3} to 10^{-3}. The corresponding Spearman correlations are 0.9762, 0.9048, and 0.9143, with a mean of 0.932. These results indicate that the candidate ordering is stable under the tested threefold change in intervention scale.

The clipping threshold does not affect the results over the tested range. Candidate gradient norms remain below the smallest threshold, $c = 5$, so clipping is never activated for $c \in \{5, 10, 20\}$. At a fixed step size, all three thresholds therefore produce identical attribution scores. The primary choice of $c = 10$ is consequently not sensitive to fine tuning of the clipping threshold.

Relation to target-encoder sensitivity. IRIS has no EMA target encoder analogous to the V-JEPA 2 target pathway, so there is no corresponding EMA setting to vary in this architecture. Target-encoder coupling sensitivity for V-JEPA 2 is evaluated separately in Appendix C.2. Taken together, these experiments indicate that the primary IRIS result does not depend on a narrowly chosen forecasting objective or intervention configuration. Restricted forecasting objectives preserve the highest-impact region, the ranking remains stable over the tested intervention scales, and clipping is inactive throughout the evaluated range.

C.4 Parameter-Subspace Sensitivity

The primary V-JEPA 2 configuration applies the forgetting-directed intervention to the final context-encoder block, its normalization layer, and the predictor. We evaluate how sensitive the attribution ranking is to this choice of parameter subspace. Using $N = 1000$ candidates and $Q = 20$ fixed reference queries, we compare the default configuration with three alternatives: an encoder-only intervention, a predictor-only intervention, and a wider subspace that includes the final two context-encoder blocks, the normalization layer, and the predictor.

Figure 12(a) shows that the ranking is stable when the intervention subspace is expanded slightly beyond the default configuration. Extending the update from the final context-encoder block to the final two blocks gives an overall Spearman correlation of $\rho = 0.9999$ with the default ranking and retains all Top-100 candidates. Restricting the intervention to only the encoder or only the predictor changes the ranking more substantially, with overall correlations of 0.708 and 0.806, respectively.

The component-level results in Figure 12(b) show where these differences arise. With encoder-only updates, the representation and structural rankings remain nearly unchanged ($\rho_R = 0.9996$ and $\rho_S = 1.0000$), while the predictive ranking changes substantially ($\rho_P = 0.0687$). Predictor-only updates, in contrast, preserve the predictive ranking ($\rho_P = 0.9917$). Because the context encoder is

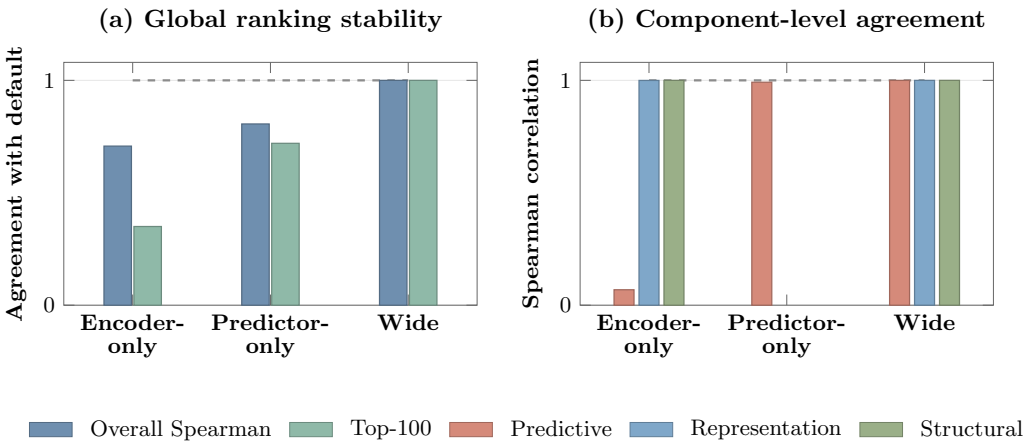

Figure 12: Parameter-subspace sensitivity of WorldTraces relative to the default V-JEPA 2 intervention subspace. (a) Overall ranking agreement and Top-100 overlap. (b) Component-level Spearman correlation. Representation and structural correlations are omitted for predictor-only because the encoder is fixed.

fixed in this setting, the representation and structural responses are constant across candidates, so their rank correlations are undefined. The wider subspace preserves all three component rankings ($\rho_P = 0.9998$, $\rho_R = 0.9999$, and $\rho_S \approx 1.0000$).

These results show that the default ranking is not sensitive to a small change in the boundary of the intervention subspace. Adding one earlier context-encoder block leaves both the global and component-level rankings essentially unchanged. Restricting the intervention to a single model pathway has a larger effect because it limits which behavioral responses can change, consistent with the predictive, representational, and structural readouts used by WorldTraces.

C.5 Readout-Definition Sensitivity

The representation and structural responses in Section 3.2 depend on particular readout definitions. We test how these choices affect the WorldTraces ranking using $N = 1{,}000$ candidates and $Q = 20$ fixed reference queries. We hold the pretrained checkpoint, candidate and query sets, forgetting-directed intervention, attribution parameter subspace, predictive response, and aggregation rule fixed, and vary only the representation or structural readout.

For representation consistency, we replace the canonical transformation $\mathcal{T}$ with either a 7×7 Gaussian blur ($\sigma = 1.5$) or a 16-pixel spatial translation down and to the right with reflected boundaries. For temporal structure, we replace the canonical first-order transition with either a second-order temporal difference or a temporal Gram descriptor constructed from pairwise cosine similarities among normalized pooled temporal states. To isolate the effect of each readout change, we substitute only the corresponding raw behavioral response and retain the other two canonical responses. We then recompute the tie-aware query-wise ranks and the fixed 1:1:1 WorldTraces aggregation.

Figure 13(a) shows that the global ranking remains similar across all four readout alternatives. Gaussian blur and spatial translation yield Spearman correlations of 0.977 and 0.978 with the canonical ranking. Their Top-100 overlaps are 91.0% and 87.0%, and their Top-500 overlaps are 94.2% and 94.0%, respectively. The representation component is much more sensitive to the transformation choice. As shown in Figure 13(b), its correlation with the canonical component ranking is approximately zero under Gaussian blur and 0.184 under spatial translation. The global ranking changes much less, showing that changes in the representation-specific ordering have a limited effect after aggregation with the predictive and structural components.

The structural alternatives show the same distinction between component-level and global ranking sensitivity, although to different degrees. Replacing the first-order transition with a second-order temporal difference changes little in either the structural component or the global ranking, with

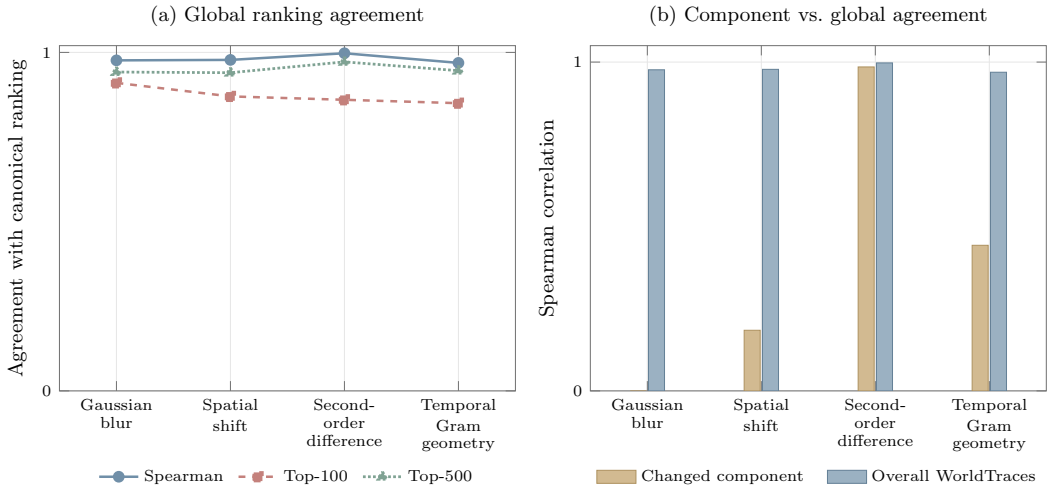

Figure 13: Readout-definition sensitivity of WorldTraces with $N = 1{,}000$ candidates and $Q = 20$ fixed reference queries. Panel (a) reports agreement between each alternative global ranking and the canonical WorldTraces ranking using Spearman correlation, Top-100 overlap, and Top-500 overlap. Panel (b) uses paired bars to compare the Spearman agreement of the changed behavioral component with that of the resulting global WorldTraces ranking. Higher values indicate stronger agreement with the canonical ranking.

Spearman correlations of 0.985 and 0.997, respectively. The corresponding Top-100 and Top-500 overlaps are 86.0% and 97.2%. The temporal Gram descriptor produces a larger change in the structural component, whose correlation with the canonical structural ranking drops to 0.443. The global WorldTraces ranking remains much closer to the canonical ranking, with a Spearman correlation of 0.969, a Top-100 overlap of 85.0%, and a Top-500 overlap of 94.6%.

Across the four alternatives, global Spearman correlations range from 0.969 to 0.997, and Top-500 overlap ranges from 94.0% to 97.2%. Individual behavioral components can therefore depend strongly on how representation consistency or temporal structure is defined, while the aggregated WorldTraces ranking is less sensitive to any single readout choice. This result supports the predefined multi-behavior aggregation and also clarifies the scope of the attribution: WorldTraces measures training data with respect to specified behavioral readouts rather than defining a readout-independent notion of data importance.

D NOTATION

Table 2 summarizes the core notation used in the WorldTraces formulation and method.

Table 2: Core notation for WorldTraces.

Symbol	Meaning
Data, model, and attribution units	
$D,\ v$	Pretraining dataset and an individual video, respectively.
Θ_D	V-JEPA 2 model pretrained on D.
$Z = \{z_i\}_{i=1}^{N},\ z$	Candidate pool of size N; each candidate z is a video or a fixed data shard, and $\lvert z\rvert$ denotes its number of videos.
$q,\ Q$	Reference query video and the fixed query set used to aggregate query-conditioned attribution scores.
$k \in \{P, R, S\}$	Behavioral component index for predictive, representational, or structural response.
$F_k(q; \Theta)$	Behavior-specific readout of model Θ on query q.
d_k	Change measure associated with behavioral component k.
$a_k^\star(z, q)$	Conceptual leave-one-out behavioral influence of candidate z on query q.
Forgetting-directed checkpoint intervention	
$E_\theta,\ P_\phi,\ E_{\bar\theta}$	Context encoder, predictor, and EMA target encoder, respectively.
$\theta,\ \phi,\ \bar\theta$	Parameters of the context encoder, predictor, and EMA target encoder, respectively.
$\omega = (\theta, \phi)$	Parameters of the online context encoder and predictor.
$\Omega,\ \Pi_\Omega$	Fixed intervention parameter subspace and projection operator that masks parameter coordinates outside Ω.
$\mathcal{L}_{\mathrm{JEPA}}(v; \omega, \bar\theta)$	V-JEPA 2 latent-prediction loss for video v.
g_z	Gradient of the average JEPA loss over candidate z, projected onto the intervention subspace Ω.
$c,\ \mathrm{clip}_c(g)$	Clipping threshold and norm-clipping operator, $\mathrm{clip}_c(g) = g/\max(1, \lVert g\rVert_2/c)$.
η	Step size of the candidate-specific forgetting-directed intervention.
$\omega_z^+,\ \theta_z^+,\ \Theta_z^+$	Intervened online parameters, their context-encoder component, and the resulting temporary checkpoint $\Theta_z^+ = (\omega_z^+, \bar\theta)$.
Behavioral responses	
$M_{\mathrm{past}},\ M_{\mathrm{future}}$	Observed prefix and held-out suffix used by the predictive attribution probe.
$\mathcal{L}_{\mathrm{future}}(q; \omega, \bar\theta)$	V-JEPA 2 latent-prediction loss evaluated on the held-out future suffix of query q.
$[x]_+$	Positive-part operator, defined as $\max(x, 0)$.
$I_P(z, q)$	Predictive response: positive part of the change in future-latent prediction loss after the candidate intervention.
$\mathcal{T}$	Fixed semantics-preserving transformation used in the representation readout.
N_q	Number of full-view context-encoder tokens for query q.
$r_\theta(q)$	Pooled and unit-normalized context-encoder representation of query q.
$V_R(q; \theta)$	Cross-view representation consistency between q and $\mathcal{T}(q)$.
$I_R(z, q)$	Representation response: unsigned change in cross-view representation consistency after intervention.
$N_t,\ N_s,\ d$	Numbers of temporal positions and spatial positions, and the latent-state dimensionality, respectively.
$H_{t,s}^\theta(q)$	Full-view encoder token at temporal position t and spatial position s.
$h_t^\theta(q)$	Spatially pooled latent state at temporal position t.
$\delta_t^\theta(q)$	First-order temporal transition, $\delta_t^\theta(q) = h_{t+1}^\theta(q) - h_t^\theta(q)$.
$I_S(z, q)$	Structural response: mean squared change in successive latent transitions after intervention.
Ranking and aggregation	
$r_{k,q}(z)$	Descending rank of candidate z for behavioral component k and query q, with average ranks assigned to ties.
$\widetilde{I}_k(z, q)$	Query-wise normalized component rank, scaled to $[0, 1]$.
$\mathrm{WorldTraces}(z, q)$	Query-conditioned WorldTraces score obtained by taking the fixed equal-weight arithmetic mean of the three normalized component ranks.
$\mathrm{WorldTraces}(z)$	Global candidate score obtained by averaging $\mathrm{WorldTraces}(z, q)$ over the fixed query set Q.
$\mathcal{S},\ K = \lvert\mathcal{S}\rvert$	Selected candidate subset and its size in a grouped intervention.