WORLDEFFECT: LEARNING ACTION-CONDITIONED OBJECT DYNAMICS FROM HUMAN PLANS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing world models primarily predict future observations, but embodied systems must additionally infer how a planned human action changes the physical state of objects in the environment. This problem becomes difficult when nearly identical motions lead to different object responses, as when a hand touches an object or narrowly misses it. Sparse contact can be obscured by shared scene motion, the coordinates that best describe an interaction change with the grasp, and even a specified plan can admit multiple outcomes. We introduce **WorldEffect**, an action-conditioned object-dynamics framework that predicts object responses from observed history and a supplied human plan. WorldEffect learns which aspects of a planned action induce object changes by comparing responses to plans with the same history. It combines coordinate views that separately assess translation and rotation, then assigns plan-conditioned probabilities to a fixed set of plausible outcomes without seeing the target response. On the BEHAVE dataset, WorldEffect reduces mean object translation error by 16.01% over fixed contact transport, which moves the object according to the planned human motion. For the same 16 candidate trajectories, its input-conditioned ranking reduces probability-weighted joint trajectory error by 12.73% relative to uniform weighting. Together, these results show that explicitly modeling action-induced object responses improves both deterministic prediction and uncertainty-aware future ordering.

1 INTRODUCTION

Video forecasting describes possible observations; embodied systems must also infer how a chosen human motion changes object states. Generative environments increasingly expose latent or explicit action interfaces (Bruce et al., 2024; Hu et al., 2023; Chen et al., 2024; Wu et al., 2024; Huang et al., 2026), while egocentric simulators condition visual futures on camera motion, body pose, or hand geometry (Tu et al., 2025; Kim et al., 2026; Wang et al., 2026; Li et al., 2026). However, these forecasts do not explicitly attribute object-state changes to the selected human action when similar visual histories precede different planned motions and object-state consequences.

We study **action-conditioned structured world dynamics**, defined here as predicting object-state transitions induced by a supplied human motion plan rather than extrapolating visual futures. We refer to the predicted object-state sequence as the object response, and use paired response differences under shared histories to analyze action effects. A suitcase should rise when the hand lifts its handle, but stay still when the hand passes by. The model takes a structured observed state and the complete supplied human plan, not raw RGB frames; the state may be reconstructed from RGB video. It predicts short-horizon six-degree-of-freedom (6DoF) rigid poses or native center/point trajectories. Recorded motion provides oracle plans; an upstream predictor provides generated plans. Figure 1 shows alternative actions alongside performance-selected suitcase and monitor cases; the latter compare frozen Top-1 predictions with recorded trajectories by average displacement error (ADE).

However, supplying a plan does not ensure that a predictor captures its effect. Two plans can share history yet differ at a decisive contact (Figure 1a): one hand displaces an object while the other narrowly misses. Sparse contact makes this distinction easy to overlook, while image-based scores can be dominated by unchanged background (Wang et al., 2004; Zhang et al., 2018; Unterthiner et al., 2018). Grasp, visibility, and motion also change which coordinates reveal the response, with different evidence governing translation and rotation. Even then, multiple response strengths or

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.

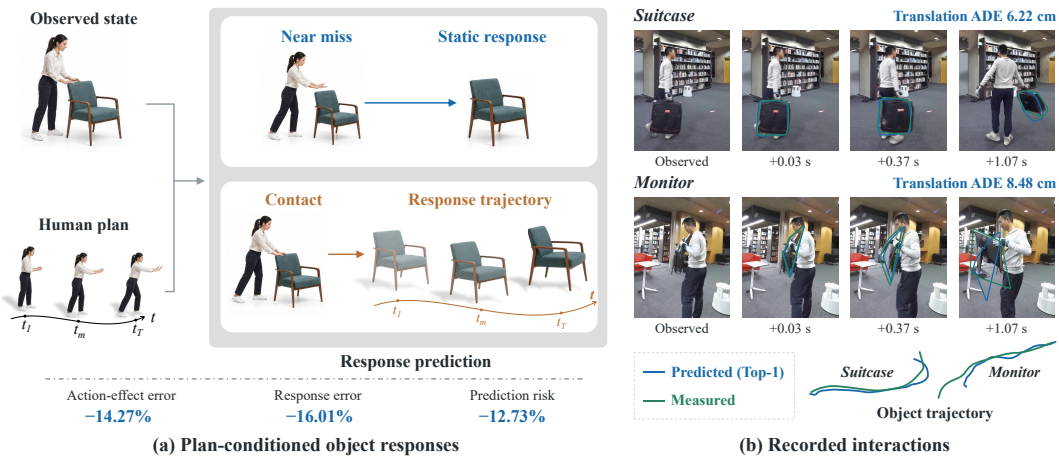


Figure 1: **Human plans shape object responses.** (a) Shared-state action-effect illustration: near-miss actions preserve the static state, while contact actions induce response trajectories. Reductions compare paired versus state-only supervision (14.27%), temporal trust versus fixed transport (16.01%), and ranking versus uniform weights (12.73%). (b) Recorded *suitcase* and *monitor* interactions: Top-1 predictions (blue), measured motions (green), and 3D translation ADE.

release times remain plausible. Prediction must identify the action effect and its coordinates while preserving uncertainty over distinct plausible outcomes.

To address these challenges, we propose **WORLDEFFECT** around *plan-response compatibility*, defined as consistency between a supplied human plan and its predicted object response. This principle organizes the prediction problem around three uncertainties: shared-history touch and near-miss pairs expose action-induced differences; complementary pose charts weight translation and rotation separately; and a fixed response cover with target-blind ordering represents distinct outcomes. *Counterfactual action pairs* are a controlled diagnostic: vary the plan under fixed observed history to assess response differences.

Experiments assess response accuracy, coverage, and ordering separately. Paired supervision lowers development trajectory error by 14.27% under matched feasibility; temporal trust lowers BEHAVE translation ADE by 16.01% over fixed transport. At fixed candidate count, structured variation improves oracle coverage and input-conditioned ordering lowers weighted risk. Independent CORE4D and video-derived Something-Something V2 (SSv2) tests show 21.39% and 5.25% error reductions against matched baselines; generated-plan tests highlight hand position and orientation.

Our contributions are summarized as follows:

- We formulate **plan-response compatibility** and a paired evaluation protocol: touch and near-miss plans share observed history but differ at contact, allowing direct comparison of predicted and target response differences across the paired actions.
- We address **action-effect, coordinate, and outcome uncertainty** through paired-effect learning, separate translation and rotation chart weights, and a fixed response cover with target-blind, plan-conditioned ordering of complete candidate trajectories.
- Tests of **response accuracy, candidate coverage, and probability ordering** show lower BEHAVE error than fixed transport, improved oracle coverage over sampled generative baselines, and 12.73% lower weighted risk with an unchanged candidate count.

2 RELATED WORK

Action-conditioned dynamics. World models evolve observations or latent states under action (Ha & Schmidhuber, 2018; Hafner et al., 2025; Wu et al., 2023a). Video and object-centric models extend this principle to controllable visual and entity dynamics (Bruce et al., 2024; Chen

et al., 2024; Wu et al., 2024; Huang et al., 2026; Agarwal et al., 2025; Guo et al., 2026; Zhao et al., 2026; Wu et al., 2023b; Zhou et al., 2025). Our structured interface measures how minimally different human plans change an object's response under a shared observed history.

Egocentric human–world interaction. Ego-view simulators connect body motion, hand geometry, and scene state (Tu et al., 2025; Kim et al., 2026; Wang et al., 2026; Li et al., 2026; Hao et al., 2026; Lin et al., 2026), while ego–exo corpora support forecasting (Grauman et al., 2022; 2024; Huang et al., 2024; Liu et al., 2022). EgoAllo and HaWoR expose complementary reference-frame invariances (Yi et al., 2025; Zhang et al., 2025); interaction corpora supply geometric supervision (Jiang et al., 2024; Bhatnagar et al., 2022; Liu et al., 2025; Chen et al., 2026; Lu et al., 2025; Kogashi et al., 2026). We address interaction-frame uncertainty in object prediction.

Structured multimodal forecasting and evaluation. Motion forecasting uses anchors, covers, coordinate-specific predictors, or intention queries (Chai et al., 2020; Phan-Minh et al., 2020; Shi et al., 2022; Zhu et al., 2023); human and hand–object generation adds goal, grasp, or physical constraints (Diomataris et al., 2024; Karunratanakul et al., 2023; Yuan et al., 2023; Li et al., 2025). HO-GCN (a human–object graph convolutional network) (Wan et al., 2022) predicts joint interaction from history and object descriptors; InterDiff (Xu et al., 2023) combines joint diffusion with contact-frame correction. Controllable Human-Object Interaction Synthesis (CHOIS) (Li et al., 2024) generates human and object motion from language, initial states, and object waypoints, whereas Object-Foresight (Soraki et al., 2026) predicts object poses from passive video. We instead fix the supplied human plan and predict its object response. Contact-relative prediction is shared with InterDiff; our mechanisms distinguish paired response differences, component-specific chart compatibility, and input-conditioned ordering of a fixed response cover.

3 METHOD

We predict object responses to supplied human plans with three linked mechanisms (Figure 2). Paired response differences isolate action effects (Section 3.2); component-specific chart weighting addresses coordinate uncertainty (Section 3.3); and a fixed response cover with target-blind, plan-conditioned ordering represents outcome uncertainty (Section 3.4).

3.1 STRUCTURED ACTION-RESPONSE FORMULATION

To separate response prediction from the choice of action, we condition on a supplied plan. Let h denote observed world history ending at $t = 0$ and $a_{1:T}$ the human motion over T future steps. We predict the resulting structured object-state transition as

$$\bar{s}_{1:T} = f_\theta(h, a_{1:T}), \tag{1}$$

where θ denotes predictor parameters. For rigid objects, s_t describes translation and rotation relative to the observed boundary pose $\boldsymbol{T}_0^o$ in the special Euclidean group $\mathrm{SE}(3)$; native adapters retain their state representation, including object centers in MMHOI and 2D point trajectories in SSv2. Plans contain body, root, or wrist trajectories and any explicitly supplied semantic conditions; object target states provide supervision and evaluation targets. Implementations with event supervision additionally predict q_t: geometric touch in TRUMANS, but a response event in BEHAVE (cumulative translation above 2 cm or rotation above 5°). These event probabilities differ from transport trust α_t and candidate weights w_k assigned to complete response trajectories at inference.

Conditioning distinguishes different actions under the same history. Passive forecasting marginalizes $p(s \mid h, a)$ over unknown actions; we retain the specified action.

In Figure 2, f_c, f_w, f_f denote contact, wrist, and factorized chart ensembles; the pose sequence and $d_{t,z}$ denote their mean pose trajectory and dispersion. "Chart prior" is the marginalized pose $\widehat{T}^{\mathrm{chart}}_{1:t}$ supplied to correction, distinct from chart weights π^0_z. *Appendix A* details the event and response pathways, feasibility, analytic chart energies, and target-blind ordering.

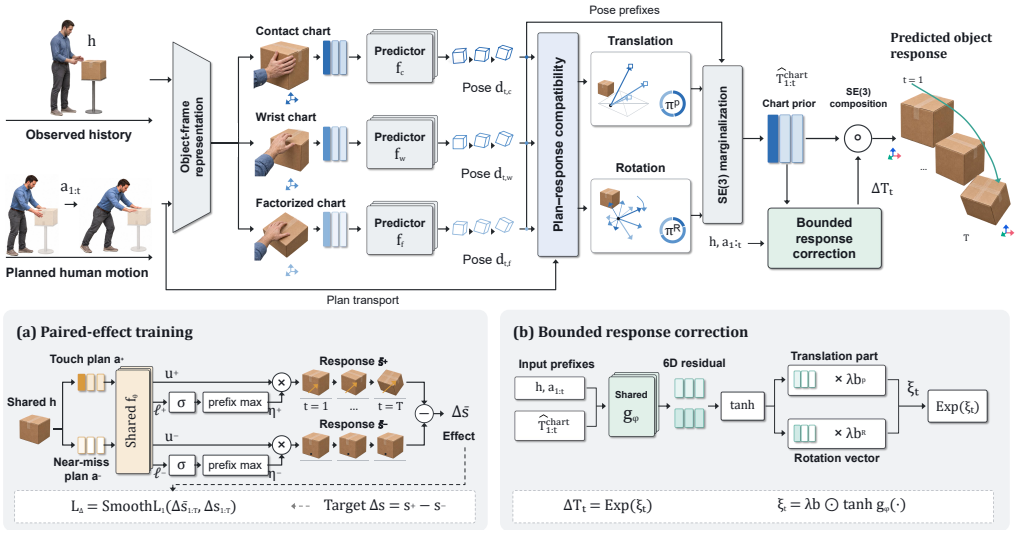


Figure 2: **Overview of WORLDEFFECT.** History h and planned motion $a_{1:t}$ define object-frame inputs. Contact, wrist, and factorized charts predict pose prefixes; plan-response compatibility determines translation and rotation weights. Marginalization gives a chart prior, followed by bounded response correction and pose composition. (a) Paired-effect training with temporal prefix event weighting. (b) Bounded translation and rotation residual correction.

3.2 IDENTIFYING ACTION-INDUCED EFFECTS

State accuracy alone need not isolate an action's effect. We therefore compare plans sharing h and camera motion: a^+ touches the object and a^- is a constructed near miss. Shared predictor weights make their response difference directly comparable. The recurrent predictor outputs unconstrained response u_t^b and event logit ℓ_t^b for each branch $b \in \{+, -\}$:

$$q_t^b = \sigma(\ell_t^b), \quad \eta_t^b = \max_{\tau \le t} q_\tau^b, \quad \bar{s}_t^b = \eta_t^b u_t^b, \quad \Delta\bar{s}_t = \bar{s}_t^+ - \bar{s}_t^-, \tag{2}$$

where σ is logistic sigmoid and $\bar{s}_t^b = [\bar{p}_t^b, \bar{r}_t^b]$ uses translation and axis-angle coordinates. The cumulative soft weight η_t^b retains established event evidence within each plan. To compare effects in interaction coordinates, human motion uses the boundary-object frame and camera motion its own boundary frame. Features encode bilateral pose, velocity, hand–object geometry, context, and time, retaining the supplied plan's native body or wrist representation.

Contact need not move an object, while an established grip can sustain transport as proximity weakens. Separate event and state heads therefore learn activation evidence and response magnitude. To learn both branch responses and their action-dependent difference, we optimize

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{\text{state}} + \mathcal{L}_\Delta + 0.5\mathcal{L}_{\text{event}} + 0.25\mathcal{L}_{\text{velocity}} + 0.5\mathcal{L}_{\text{terminal}}, \\ \mathcal{L}_\Delta &= \text{SmoothL1}(\Delta\bar{s}_{1:T}, \Delta s_{1:T}), \qquad \Delta\bar{s} = \bar{s}^+ - \bar{s}^-, \quad \Delta s = s^+ - s^-, \end{aligned} \tag{3}$$

where $\mathcal{L}_{\text{event}}$ is binary cross-entropy (BCE) on recurrent event logits. State, difference, velocity, and terminal losses supervise the soft-weighted response $\bar{s}$, as illustrated by the paired objective in Figure 2(a). These soft-weighted responses supply the chart fusion and candidate-ordering stages.

3.3 COORDINATE-AWARE OBJECT RESPONSE

Object-relative coordinates expose hand–object proximity; wrist-relative coordinates expose motion from the observed wrist. Their reliability varies with contact and visibility. We encode hand history and plan in three input charts $z \in \mathcal{Z} = \{c, w, f\}$ (Figure 2): contact (c) uses the boundary-object frame; wrist (w) uses each hand's boundary-wrist frame; factorized (f) combines contact-frame history with wrist-anchored planned motion. Each chart feeds an expert predictor; all share architecture and training target, directly predicting boundary-object displacement and relative rotation. The predicted responses $\bar{s}_t^z$ have within-chart dispersion $d_{t,z}$.

To weight translation and rotation separately, we compare each expert with motion $(\Delta h_t, \Delta R_t^h)$ implied by visible hand-plan anchors and define standardized component disagreement

$$c_{t,z}^p = \mathrm{Z}_{\mathcal{Z}}(\|\bar{p}_t^z - \Delta h_t\|_2)\,, \qquad c_{t,z}^R = \mathrm{Z}_{\mathcal{Z}}\big(d_{\mathrm{geo}}(\bar{R}_t^z, \Delta R_t^h)\big)\,. \tag{4}$$

To favor reliable charts whose motion agrees with the plan, we combine this disagreement with ensemble dispersion in separate component energies:

$$e_{t,z}^p = \mathrm{Z}_{\mathcal{Z}}(d_{t,z}) + \alpha g_t c_{t,z}^p, \quad e_{t,z}^R = \mathrm{Z}_{\mathcal{Z}}(d_{t,z}) + \alpha g_t (c_{t,z}^p + \beta \kappa_t c_{t,z}^R), \quad \widetilde{\pi}_{t,z}^u \propto \pi_z^0 e^{-e_{t,z}^u}. \tag{5}$$

For component $u \in \{p, R\}$, normalized weights $\widetilde{\pi}^u$ are mixed with the fixed chart prior π^0 as $\pi^u = \varepsilon_{\mathrm{prior}}\pi^0 + (1 - \varepsilon_{\mathrm{prior}})\widetilde{\pi}^u$, with $\varepsilon_{\mathrm{prior}} = 0.4$. α and β weight plan disagreement and its rotational component; d_{geo} is the rotation geodesic distance, and $\bar{R}_t^z$ represents the chart's predicted rotation matrix. Here $\mathrm{Z}_{\mathcal{Z}}$ standardizes each mismatch separately across charts at fixed sample and time. The prior floor preserves support for each chart; numerical standardization and reliability definitions appear in *Appendix A.3*. Component-wise marginalization yields

$$\hat{p}_t = \sum_z \pi_{t,z}^p \bar{p}_t^z, \quad \hat{R}_t = \rho_6\left(\sum_z \pi_{t,z}^R \bar{u}_t^z\right), \tag{6}$$

where ρ_6 orthonormalizes the weighted first two rotation columns $\bar{u}_t^z \in \mathbb{R}^6$. The gate g_t suppresses compatibility when neither hand approaches; rotation reliability κ_t decreases under weak rotation or bilateral disagreement. Translation and rotation can thus favor different charts without a learned router. Observed anchors keep the reference geometry fixed while the supplied hand motion determines transport. HOT3D tests this geometric agreement in calibrated pose prediction.

To adapt the geometric prior to interaction-specific response errors, we apply shared, identity-free heads for *bounded response correction* (Figure 2(b)):

$$\begin{aligned} \delta p_t &= b^p \tanh g_\phi^p(\cdot), \quad & p_t^{\mathrm{resp}} &= p_t^{\mathrm{chart}} + \delta p_t, \\ \delta \omega_t &= b^R \tanh g_\phi^R(\cdot), \quad & R_t^{\mathrm{resp}} &= R_t^{\mathrm{chart}} \exp([\delta\omega_t]_\times). \end{aligned} \tag{7}$$

The heads g_ϕ^p, g_ϕ^R have learned parameters ϕ; $[\cdot]_\times$ maps a rotation vector to its skew-symmetric matrix and exp is the matrix exponential. History, plan prefixes, and chart predictions condition componentwise corrections bounded by $b^p = 6$ cm and $b^R = 12^\circ$. Figure 2(b) collects them as ξ_t: its $\mathrm{Exp}(\xi_t)$ is shorthand for the separate translation update and special orthogonal group $\mathrm{SO}(3)$ rotation exponential in Equation 7; $\odot$ denotes elementwise multiplication. The panel sketches attenuation; *Appendix A.3* gives the exact translation contraction and rotation blend used for population evaluation. Full-body manipulation requires balancing learned motion with persistent contact transport as the interaction develops. In BEHAVE, temporal trust resolves this choice by adaptively combining learned motion u_t with transport v_t from persistent contact anchors:

$$\bar{s}_t^j = E_t\big[(1 - \alpha_t^j)u_t^j + \alpha_t^j v_t^j\big], \qquad E_t = \max_{\ell \le t} q_\ell, \quad j \in \{p, r\}. \tag{8}$$

The sigmoid weights $\alpha_t^p, \alpha_t^r \in [0, 1]$ use plan geometry, transport disagreement, and event statistics; r is an axis-angle vector. Fixed transport uses constant weights, whereas temporal trust adapts their contribution at each step. *Appendix A.3* specifies the boundary geometry, complete-plan features, and transformation of outputs back to the world frame.

3.4 Multimodal Futures and Target-Blind Ordering

Figure 2 shows the single-response path; a fixed set of response parameters generates complete alternatives, which plan evidence then orders. Coordinate choice leaves ambiguity in response strength, release time, and transport source. Rather than average away these alternatives, we separate their representation from probability assignment. We vary translation/rotation response and temporal trust to form candidate parameters. Alternating average/final displacement error (ADE/FDE) facility location selects $K = 16$ codewords on sequence-disjoint development. This cover is a vector quantization (VQ) of response parameters. Each codeword (s_p, s_R, o_p, o_R) replaces α_t^j in Equation 8 by $\mathrm{clip}(\alpha_t^j + o_j, 0, 1)$ and scales the resulting component by s_j. Offsets modify trust weights directly; scaled rotation vectors are exponentiated to valid rotations. The parameters are fixed, but

the complete candidate trajectories depend on (h, a). *Appendix B.3* specifies pool construction and selection. Let $S_K(h, a)$ denote these trajectories. Coverage alone does not tell us which response the input supports, so we learn preference logits $r_\psi(x)_k$ from plan, event, and contact evidence $x = x(h, a)$. To favor low-risk candidates while retaining alternatives, training minimizes

$$\mathcal{L}_{\text{rank}} = \sum_{k=1}^{K} p_k R_k + \lambda_{\text{ce}} \, \text{CE}(p, k^*) - \lambda_H H(p), \qquad p = \text{softmax}(r_\psi(x)), \qquad w = \text{softmax}(\tau r_\psi(x)), \tag{9}$$

where $R_k = J_k / \max(K^{-1} \sum_j J_j, 10^{-6})$ normalizes joint error J_k, $k^* = \arg\min_k J_k$, $\text{CE}(p, k^*) = -\log p_{k^*}$, and $H(p) = -\sum_k p_k \log p_k$. Training uses p; inference uses w. The loss weights $\lambda_{\text{ce}}, \lambda_H$ and τ (an inverse-temperature multiplier) are frozen using development data; *Appendix A.2* gives their values. The parameters ψ define the preference network. At test time the ranker sees complete-plan statistics x but neither R_k nor k^*; it only assigns a distribution over the frozen cover. Deployable ordering and oracle coverage measure distinct capabilities.

4 Experiments

4.1 Experimental Setup

Datasets and prediction contract. We use TRUMANS (Jiang et al., 2024) pairs to study action effects, HOT3D (Banerjee et al., 2025) geometry to test coordinate reasoning, and BEHAVE (Bhatnagar et al., 2022) full-body interactions to evaluate response uncertainty (9,952/3,887 train/test windows). Further interaction tests use HUMOTO, MMHOI, CORE4D, and HOI-M3 (Lu et al., 2025; Kogashi et al., 2026; Liu et al., 2025; Zhang et al., 2024). SSv2 tests video-derived states (Goyal et al., 2017; Mahdisoltani et al., 2018); HO-GCN tests plan noise (Wan et al., 2022). Splits precede window extraction. Plans use recorded human motion unless forecast; object trajectories remain targets. RGB supplies offline state extraction and visualization rather than dynamics inputs.

Comparisons, metrics, and statistics. To distinguish action use from motion extrapolation, we compare zero motion, proximity, and history-only controls, then perturb the supplied plan. Gated recurrent units (GRUs), residual models, and budget-matched plan Transformers test learned alternatives. For a single trajectory, ADE averages error over time and FDE measures endpoint error; translation uses centimeters and rotation uses geodesic degrees. For multiple futures, oracle error measures coverage, while energy score (ES) evaluates the weighted translation distribution; lower is better for both. Image-space tests measure projected position error and intersection over union (IoU, higher is better). In Table 1, bold and underlined values mark the best and second-best distinct scores, respectively; tied scores share a rank. Confidence intervals (CIs) resample independent sequences or pairs. The seven TRUMANS pairs also select feasibility, so their results measure development performance rather than independent confirmation.

4.2 Action-Induced Effect Prediction

We first compare the complete WorldEffect method with matched-input predictors on fourteen diagnostic pairs from three scenes (Table 1). WorldEffect combines paired learning, chart fusion, and candidate ranking on soft-weighted responses; the learned alternatives receive the same geometric views and planned transport. Single-output selection minimizes distance to the predicted distribution, without reading the target. WorldEffect leads on endpoint translation and rotation errors; the learned set has lower translation ADE and ES. Figure 3 separately shows the value of complete plans in MMHOI. *Appendix B.2* details these matched controls.

We next isolate action effects with TRUMANS pairs sharing history and camera motion. Only the planned right-arm path changes from recorded touch to collision-free near miss, whose object target is analytically static rather than a separately executed physical outcome. For this constructed-pair diagnostic only, an external, target-free binary mask uses planned wrist geometry and event confidence: $\hat{s}_t^b = M_t^b \bar{s}_t^b$ after soft weighting. It is unused elsewhere; *Appendix B.2* gives its frozen rule. Table 2 reports Pair ADE across both branches, touch FDE, and near-miss ADE to separate response accuracy from unsupported motion on the constructed near-miss branch.

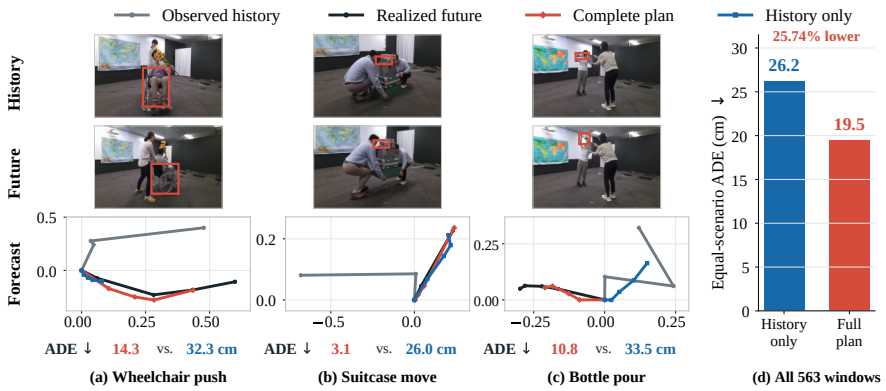


Figure 3: **Human-plan information changes the response.** (a–c) Complete-plan WorldEffect (red) and history-only control (blue) on three preregistered MMHOI actions. (d) Equal-scenario ADE summarizes all 563 validation windows across the evaluated scenarios.

Table 1: **Action-conditioned prediction.** WorldEffect has the lowest endpoint errors on 14 diagnostic pairs. T/ES: cm; R: degrees; A/F: ADE/FDE.

Method	K	T-A↓	T-F↓	R-A↓	R-F↓	ES↓
Zero motion	1	2.528	10.762	4.685	17.240	4.233
Plan GRU	1	3.539	10.195	6.779	20.247	5.237
Plan Transformer	1	3.900	10.526	6.438	18.280	5.060
Learned set	16	**2.224**	8.615	6.305	20.831	**2.780**
WorldEffect	16	2.273	**8.439**	**4.529**	**16.285**	3.133

Table 2: **Controlled action effects.** Paired loss lowers Pair ADE. Pair/near: ADE; touch: FDE (cm).

Method	Pair↓	Touch↓	Near↓
Zero motion	1.876	18.485	0.000
Proximity	2.893	17.197	1.355
State + gate	1.641	11.909	0.000
Transformer	1.653	13.590	0.000
WorldEffect	1.407	11.516	0.000

Effect identification and temporal factorization. WorldEffect reduces Pair ADE by 25.01% relative to zero motion and 14.88% relative to the matched Transformer. The external binary mask enforces zero drift on constructed static near misses. Holding soft weighting fixed, we vary the binary mask and paired loss: paired supervision improves ADE by 14.27% under the same mask, and masking adds 7.05% with the paired objective fixed. Learning the response difference and suppressing unsupported motion therefore make distinct contributions.

Recorded and executed action alternatives. We extend this evidence beyond constructed touch/near-miss contrasts using recorded interaction pairs and same-state physics executions. On 22 approximately history-matched BEHAVE recording pairs, WorldEffect reduces measured response-difference ADE from the plan Transformer's 5.90 to 5.63 cm (*Appendix B.2*).

To execute both alternatives from exactly the same state, we additionally simulate a prescribed spherical hand proxy pushing, bypassing, or stopping before a rigid object. On 96 held-out initial configurations, plan conditioning lowers six-second response-difference ADE from history-only's 13.305 to 0.839 cm. Ordinary state supervision reaches 0.842 cm, so this controlled physics test establishes action dependence without a clear six-second advantage for the paired loss (*Appendix B.2*).

The contribution of supplied human plans. On three preregistered MMHOI actions and all 563 validation windows, complete plans lower scenario-averaged ADE from 26.20 to 19.46 cm (25.74%). Thus the supplied action contributes information beyond history extrapolation, consistent with the controlled action-effect comparison under matched initial conditions.

4.3 Object Response Prediction

On the official BEHAVE split, contact can persist or release, and no matched near miss is available. We project predicted object poses into calibrated cameras to assess visible placement (Figure 4).

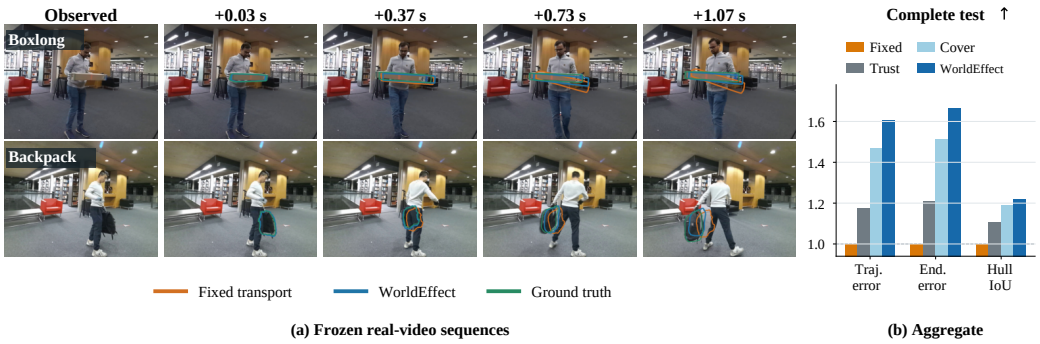


Figure 4: **Structured gains carry into camera space.** (a) Fixed cases compare transport (orange), a pixel-oracle trajectory (blue), and measured poses (green). (b) Trust, Cover, and WorldEffect improve test-set placement over fixed transport; cover bars use oracle selection.

Table 3: **Full-body response.** Trust lowers all single-response errors; the cover improves oracle coverage. BEHAVE test; T/R: cm/degrees; K: candidate count.

Method	K	T-ADE↓	T-FDE↓	R-ADE↓	R-FDE↓
Zero motion	1	16.687	28.820	22.047	38.218
Prefix Transformer (Vaswani et al., 2017)	1	9.737	18.148	17.546	31.832
Fixed contact transport	1	7.821	13.741	14.822	25.579
WorldEffect (+ trust)	1	6.569	11.137	13.547	22.645
Conditional diffusion	16	7.283	10.867	14.262	20.352
WorldEffect trust–response cover	16	4.633	6.702	10.318	15.120

Table 4: **Candidate ordering.** Learned weights reduce joint risk on BEHAVE.

Metric	Uniform	Ridge	Exp. util.
Wt. risk↓	2.688	2.437	2.346
Risk red.↑	0.00%	9.34%	12.73%
Utility 95% CI		$[11.56, 13.92]\%$	
Top-1 hit rate↑	6.25%	12.58%	22.07%
Top-4 hit rate↑	25.00%	39.44%	52.92%
ES (cm)↓	6.033	6.005	5.917
Brier↓	0.9375	0.9262	0.8937

Deterministic response. The $K = 1$ block of Table 3 compares deterministic predictors. Fixed transport improves on the plan Transformer by expressing how a sustained grip carries an object. Temporal trust then lowers all four errors by adapting transport weights as the interaction changes. Under the same supplied plans, translation ADE falls from 7.821 to 6.569 cm (16.01%).

The evaluation uses all 3,887 test windows; its qualitative panel fixes one pixel-ADE oracle across the horizon. The aggregate bars report fixed/method ratios for pixel errors and method/fixed ratios for IoU; Trust uses one response, while Cover and WorldEffect report oracle candidate coverage. Temporal trust lowers pixel trajectory error from 33.42 to 28.48 px; the cover's pixel oracle reaches 20.84 px, measuring available responses rather than deployable selection. Closer projected paths and higher outline overlap connect trajectory accuracy to object placement.

4.4 Multimodal Response Uncertainty

Coverage and probability ordering. The $K = 16$ block of Table 3 evaluates whether the candidates contain an accurate response, selecting a complete trajectory separately for each metric. Diffusion and flow generate trajectories from sampled noise; WorldEffect varies response magnitude and the temporal balance between learned response and planned transport around a shared base prediction. These structured variations express interaction-related uncertainty through response parameters. Trust–response variation improves oracle coverage over magnitude variation, learned sets, diffusion, and flow. Same-base residual flow isolates the uncertainty head: structured variation improves coverage while retaining the same deterministic response.

We next hold these trajectories fixed and compare their probability assignments (Table 4). Uniform assigns $1/16$ to each candidate. Ridge is a regularized linear regressor fitted on development data to predict candidate errors from input features; a softmax of negative predicted errors assigns higher probability to lower-risk candidates. Our expected-utility ranker instead learns probabilities with Equation 9. Weighted risk averages normalized translation and rotation ADE/FDE using these weights. Top-k hit rate asks whether the lowest-risk candidate is among the k highest-ranked; Brier measures squared probability error for that candidate's identity. Expected utility lowers weighted

Table 5: **Component analysis.** Equal chart weights and removing the cover worsen error on 14 diagnostic pairs. ADE/ES: cm; risk: normalized joint ADE; bold/underlined: best/second-best.

Configuration	T-ADE	Weighted risk	ES
WorldEffect	<u>2.273</u>	**2.205**	<u>3.133</u>
Without paired supervision	**2.260**	<u>2.229</u>	3.194
Equal chart weights	2.915	2.474	3.721
Without cover ($K = 1$)	2.588	2.481	3.770
Uniform candidate probabilities	2.576	2.466	**2.907**

Table 6: **Prediction across interaction domains.** WorldEffect improves over matched baselines on independent roles. Errors use each dataset's native metric; units are shown by row.

Dataset	Role / units	Metric	Comparator	Base↓	Ours↓	Gain↑
HUMOTO	Test / 11 seq.	Composite ADE (cm)	Constant velocity	11.66	5.85	49.80%
CORE4D	Confirmation / 24 seq.	Normalized error	Plan Transformer	0.720	0.566	21.39%
SSv2	Evaluation / 104 clusters	Point ADE (px)	Absolute-plan Transformer	34.01	32.23	5.25%

risk by 12.73% versus uniform weighting, with a CI from paired bootstrap resampling over 82 sequences. It also improves ES over Ridge and a fixed nonuniform prior (*Appendix B.3*), supporting input-dependent probability assignment. The most probable trajectory has 5.00% higher joint risk than the strongest $K = 1$ prediction. Its confidence is also less consistent with hit rate, distinguishing distribution quality from single-output accuracy alone.

Contributions of compatibility components. Table 5 removes one component from the released pipeline on the same fourteen diagnostic pairs as Table 1. Equal chart weights and a single identity candidate increase all three errors, supporting coordinate resolution and candidate coverage. The no-pair experts are retrained at the same budget; other removals retain learned experts and fixed codewords and router, apart from the single-candidate and uniform-probability controls. Paired supervision constrains response differences alongside absolute states. It lowers weighted joint risk and translation ES despite slightly higher T-ADE, linking action effects to distribution quality. Uniform probabilities lower ES but raise normalized joint risk.

4.5 Evidence Across Interaction Domains

Table 6 extends the comparisons to multi-person interactions and video-derived states. Each comparison retains its native metric and pairs predictions under the same input conditions and evaluation role, separating domain coverage from transfer of a shared checkpoint. HUMOTO's larger gain is against constant velocity, whereas the CORE4D and SSv2 controls already receive planned motion. Their gains therefore show that response modeling helps beyond plan access.

Additional imperfect- and generated-plan experiments examine how upstream plan quality affects downstream object response (*Appendix B.4*), distinguishing accurate human motion from motion that preserves interaction-relevant hand position and orientation.

5 Conclusion

We present WorldEffect, a framework for action-conditioned object dynamics that predicts structured object responses to supplied human-motion plans. Paired action-effect learning, component-specific charts, and target-blind cover ordering address action, coordinate, and outcome uncertainty. Controlled and recorded comparisons show more accurate object responses; structured candidates improve oracle coverage, and learned probabilities reduce weighted risk. Oracle coverage measures the candidate set's potential; the issued probabilities order those alternatives under the supplied plan. Together, the results distinguish accurate single responses from well-ranked alternatives over plausible futures. Cross-domain tests support this prediction contract under different state representations. Generated-plan analysis highlights interaction-relevant wrist motion. Future work explores articulated objects and feedback-driven replanning over longer human–object interactions.

AI USE STATEMENT

In this work, we use generative AI tools for literature retrieval and discovery, research ideation and execution, drafting manuscript sections, and editing language and presentation. The authors set the research questions, provide domain judgment, review designs and interpretations, and revise the manuscript. AI assistance includes identifying and summarizing related work; proposing and refining hypotheses, conceptual explanations, and experimental protocols; implementing and executing data-processing, training, evaluation, and analysis code; interpreting diagnostics; drafting and restructuring text; and creating or revising scientific figures. Literature and technical claims are checked against original sources and experimental evidence. Reported results come from executed experiments and recorded artifacts, not language-model estimates. Evaluation targets derive from recorded data and the disclosed state-reconstruction pipelines; analytically constructed near-miss targets are distinguished from measured outcomes in *Appendix A.1*. Checks on AI-assisted implementation include unit tests, input and data-provenance checks, numerical regression tests, and independent manuscript rebuilds. Human authors review AI-assisted work, retain final scientific judgment, and take responsibility for the submitted text, claims, code, figures, and results.

ETHICS STATEMENT

The authors read and adhere to the ICLR Code of Ethics. This study uses publicly released research datasets under their respective terms and does not recruit, contact, or intervene with new human participants. Some source datasets contain human motion and egocentric video, which can expose sensitive behavior and reflect demographic, geographic, capture-device, and activity biases. The method consumes task-relevant motion and interaction annotations, excludes identity labels from its response graph, and performs neither identity recognition nor inference of personal attributes. The released research artifacts contain code, protocols, and derived aggregate results rather than redistributed raw personal video; users must obtain source data from the original providers and comply with their licenses and access conditions. The intended use is research on digital-human interaction and structured world dynamics. A predicted object response must not be treated as evidence of a person's intent, capability, or future behavior, and the method is not validated for surveillance, high-stakes decisions, or safety-critical control. The scope and limitations in *Appendix C* further document rigid-object, short-horizon, privacy, and representational boundaries.

REPRODUCIBILITY STATEMENT

The task contract, model components, and information-access assumptions appear in Section 3; Section 4 reports the principal datasets, comparators, metrics, and quantitative results. The supplementary material specifies data construction and split roles (*Appendix A.1*), architecture, objectives, optimization, and seed aggregation (*Appendix A.2*), inference, compute, and reproducibility details (*Appendix A.4*), metric and aggregation definitions (*Appendix B.1*), and the complete controlled, real-interaction, cross-corpus, and egocentric protocols (*Appendices B–B.6*). *Appendix B.7* records fixed qualitative-selection protocols and extended visual evidence; performance-screened generated-plan cases are identified separately in *Appendix B.5*. The project preserves training and evaluation implementations, configuration and split manifests, figure-generation scripts, and machine-readable result records supporting the reported tables and figures.

REFERENCES

Niket Agarwal et al. Cosmos World Foundation Model Platform for Physical AI. *arXiv preprint arXiv:2501.03575*, 2025.

Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. HOT3D: Hand and Object Tracking in 3D from Egocentric Multi-View Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Bharat Lal Bhatnagar, Xianghui Xie, Ilya A. Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. BEHAVE: Dataset and Method for Tracking Human Object Interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15935–15946, 2022.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative Interactive Environments. *arXiv preprint arXiv:2402.15391*, 2024.

Yuning Chai, Benjamin Sapp, Mayank Bansal, and Dragomir Anguelov. MultiPath: Multiple Probabilistic Anchor Trajectory Hypotheses for Behavior Prediction. In *Proceedings of the Conference on Robot Learning*, volume 100, pp. 86–99, 2020.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion Forcing: Next-token Prediction Meets Full-Sequence Diffusion. In *Advances in Neural Information Processing Systems*, 2024.

Yue Chen, Xingyu Chen, Yuxuan Xue, Anpei Chen, Yuliang Xiu, and Gerard Pons-Moll. Human3R: Everyone Everywhere All at Once. In *International Conference on Learning Representations*, 2026.

Markos Diomataris, Nikos Athanasiou, Omid Taheri, Xi Wang, Otmar Hilliges, and Michael J. Black. WANDR: Intention-guided Human Motion Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 927–936, 2024.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "Something Something" Video Database for Learning and Evaluating Visual Common Sense. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

Kristen Grauman, Andrew Westbury, Eugene Byrne, et al. Ego4D: Around the World in 3,000 Hours of Egocentric Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18995–19012, 2022.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, et al. Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19383–19400, 2024.

Yanjiang Guo, Lucy Xiaoyang Shi, Jianyu Chen, and Chelsea Finn. Ctrl-World: A Controllable Generative World Model for Robot Manipulation. In *International Conference on Learning Representations*, 2026.

David Ha and Jürgen Schmidhuber. World Models. *arXiv preprint arXiv:1803.10122*, 2018.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640:647–653, 2025. doi: 10.1038/s41586-025-08744-2.

Jinkun Hao, Mingda Jia, Ruiyan Wang, Hongrui Zhu, Jiafei Cao, Xihui Liu, Ran Yi, Lizhuang Ma, Jiangmiao Pang, and Xudong Xu. EgoSim: Egocentric World Simulator for Embodied Interaction Generation. *arXiv preprint arXiv:2604.01001*, 2026.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. GAIA-1: A Generative World Model for Autonomous Driving. *arXiv preprint arXiv:2309.17080*, 2023.

Siqiao Huang, Jialong Wu, Qixing Zhou, Shangchen Miao, and Mingsheng Long. Vid2World: Crafting Video Diffusion Models to Interactive World Models. In *International Conference on Learning Representations*, 2026.

Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, and Yu Qiao. EgoExoLearn: A Dataset for Bridging Asynchronous Ego- and Exo-centric View of Procedural Activities in Real World. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22072–22086, 2024.

Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling Up Dynamic Human-Scene Interaction Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided Motion Diffusion for Controllable Human Motion Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2151–2162, 2023.

Byungjun Kim, Taeksoo Kim, Junyoung Lee, and Hanbyul Joo. Dexterous World Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Kaen Kogashi, Anoop Cherian, and Meng-Yu Jennifer Kuo. MMHOI: Modeling Complex 3D Multi-Human Multi-Object Interactions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1512–1521, 2026.

Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C. Karen Liu. Controllable Human-Object Interaction Synthesis. In *European Conference on Computer Vision*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/5720_ECCV_2024_paper.php.

Muchen Li, Sammy Christen, Chengde Wan, Yujun Cai, Renjie Liao, Leonid Sigal, and Shugao Ma. LatentHOI: On the Generalizable Hand Object Motion Generation with Latent Hand Diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17416–17425, 2025.

Yu Li, Menghan Xia, Gongye Liu, Xintao Wang, Conglang Zhang, Lei Ke, Yuxuan Lin, Ruihang Chu, Pengfei Wan, Kun Gai, and Yujiu Yang. AnchorWorld: Embodied Egocentric World Simulation with View-based Evolution Customization. *arXiv preprint arXiv:2606.07326*, 2026.

Zuyao Lin, Jianhui Zhang, Peidong Jia, Xiaoguang Zhao, Shanghang Zhang, and Xingyu Chen. World-Ego Modeling for Long-Horizon Evolution in Hybrid Embodied Tasks. *arXiv preprint arXiv:2605.19957*, 2026.

Yun Liu, Chengwen Zhang, Ruofan Xing, Bingda Tang, Bowen Yang, and Li Yi. CORE4D: A 4D Human-Object-Human Interaction Dataset for Collaborative Object REarrangement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1769–1782, 2025.

Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. HOI4D: A 4D Egocentric Dataset for Category-Level Human-Object Interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21013–21022, 2022.

Jiaxin Lu, Chun-Hao Paul Huang, Uttaran Bhattacharya, Qixing Huang, and Yi Zhou. HUMOTO: A 4D Dataset of Mocap Human Object Interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10886–10897, 2025.

Farzaneh Mahdisoltani, Guillaume Berger, Waseem Gharbieh, David J. Fleet, and Roland Memisevic. On the Effectiveness of Task Granularity for Transfer Learning. *arXiv preprint arXiv:1804.09235*, 2018.

Tung Phan-Minh, Elena Corina Grigore, Freddy A. Boulton, Oscar Beijbom, and Eric M. Wolff. CoverNet: Multimodal Behavior Prediction Using Trajectory Sets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Motion Transformer with Global Intention Localization and Local Movement Refinement. In *Advances in Neural Information Processing Systems*, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Rustin Soraki, Homanga Bharadhwaj, Ali Farhadi, and Roozbeh Mottaghi. ObjectForesight: Predicting Future 3D Object Trajectories from Human Videos. *arXiv preprint arXiv:2601.05237*, 2026. URL https://arxiv.org/abs/2601.05237.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033, 2012. doi: 10.1109/IROS.2012.6386109.

Yuanpeng Tu, Hao Luo, Xi Chen, Xiang Bai, Fan Wang, and Hengshuang Zhao. PlayerOne: Egocentric World Simulator. In *Advances in Neural Information Processing Systems*, 2025.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards Accurate Generative Models of Video: A New Metric & Challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, 2017.

Weilin Wan, Lei Yang, Lingjie Liu, Zhuoying Zhang, Ruixing Jia, Yi-King Choi, Jia Pan, Christian Theobalt, Taku Komura, and Wenping Wang. Learn to Predict How Humans Manipulate Large-Sized Objects From Interactive Motions. *IEEE Robotics and Automation Letters*, 7(2):4702–4709, 2022.

Yuxi Wang, Wenqi Ouyang, Tianyi Wei, Yi Dong, Zhiqi Shen, and Xingang Pan. Hand2World: Autoregressive Egocentric Interaction Generation via Free-Space Hand Gestures. *arXiv preprint arXiv:2602.09600*, 2026.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. iVideoGPT: Interactive VideoGPTs are Scalable World Models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. DayDreamer: World Models for Physical Robot Learning. In *Proceedings of the Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, 2023a.

Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg. SlotFormer: Unsupervised Visual Dynamics Simulation with Object-Centric Models. In *International Conference on Learning Representations*, 2023b.

Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Brent Yi, Vickie Ye, Maya Zheng, Yunqi Li, Lea Müller, Georgios Pavlakos, Yi Ma, Jitendra Malik, and Angjoo Kanazawa. Estimating Body and Hand Motion in an Ego-sensed World. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7072–7084, 2025.

Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. PhysDiff: Physics-Guided Human Motion Diffusion Model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16010–16021, 2023.

Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. HaWoR: World-Space Hand Motion Reconstruction from Egocentric Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1805–1815, 2025.

Juze Zhang, Jingyan Zhang, Zining Song, Zhanhe Shi, Chengfeng Zhao, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. HOI-M3: Capture Multiple Humans and Objects Interaction within Contextual Environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Min Zhao, Hongzhou Zhu, Bokai Yan, Zihan Zhou, Yimin Chen, Wenqiang Sun, Kaiwen Zheng, Guande He, Xiao Yang, Chongxuan Li, Fan Bao, and Jun Zhu. minWM: A Full-Stack Open-Source Framework for Real-Time Interactive Video World Models. *arXiv preprint arXiv:2605.30263*, 2026.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World Models on Pre-Trained Visual Features Enable Zero-Shot Planning. In *Proceedings of the International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79115–79135, 2025.

Yiyao Zhu, Di Luan, and Shaojie Shen. BiFF: Bi-level Future Fusion with Polyline-based Coordinate for Interactive Trajectory Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8260–8271, 2023.

WORLDEFFECT: Learning Action-Conditioned Object Dynamics from Human Plans

Appendix

Anonymous Authors

APPENDIX CONTENTS

A. Method Details and Inference 16
A.1. Task, Data, and Information Access 16
A.2. Architecture and Learning 17
A.3. Detailed Structured Inference 18
A.4. Inference and Reproducibility 20
B. Additional Experimental Results 20
B.1. Metrics and Statistical Units 20
B.2. Paired Effects and Feasibility 20
B.3. Full-Body Prediction, Coverage, and Ordering 23
B.4. Generated and Imperfect Plans 26
B.5. Video-Derived and Multi-Actor Interactions 27
B.6. Interaction Charts and Egocentric Transfer 28
B.7. Qualitative Response and Actor Attribution 29
C. Discussion and Future Directions 31

Highlights and suggested reading. For a compact path through the principal evidence, we recommend:

Method and inference.
Architecture and objectives in Section A.2; temporal trust, chart mappings, full candidate generation, and probability ordering in Section A.3.

Controlled effects and uncertainty.
Paired-loss and feasibility controls in Section B.2; coverage, issued predictions, distribution scoring, and codeword calibration in Section B.3.

Real interactions and transfer.
Video-derived and multi-actor responses in Section B.5; component compatibility and egocentric evaluation in Section B.6.

Visual and imperfect-plan evidence.
Recorded and generated plans in Section B.4; real-image projections and additional response galleries in Sections B.5 and B.7.

A METHOD DETAILS AND INFERENCE

To make the three compatibility mechanisms reproducible, we first describe their inputs and learning signals, then their inference operators. Figure 5 connects the event and response pathways, chart weighting, and candidate ordering to their respective supervision signals.

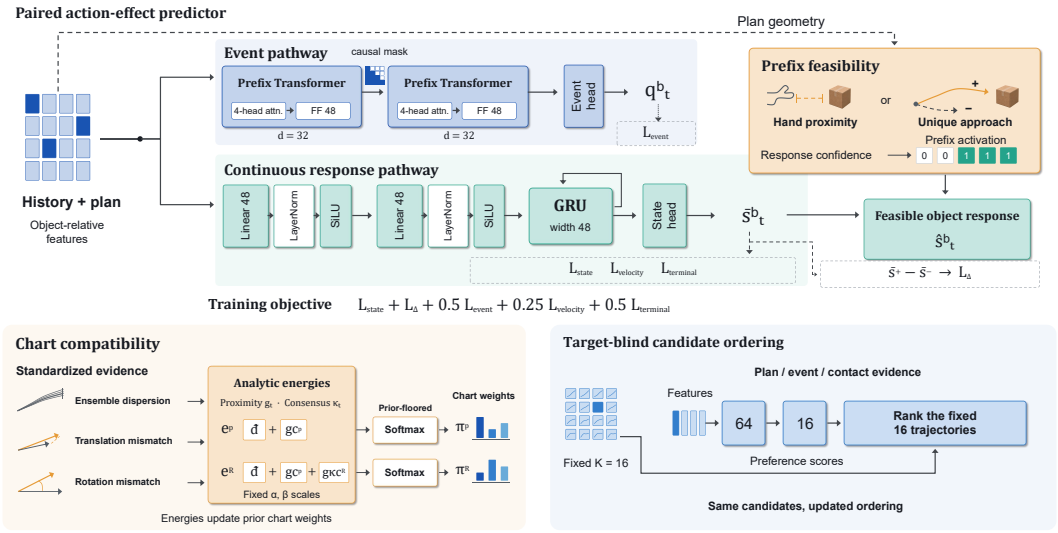


Figure 5: **Implementation details of WORLDEFFECT.** Top (TRUMANS): event and continuous response pathways, response supervision, and geometry/confidence-based feasibility. Bottom left (HOT3D): standardized dispersion and plan mismatch define analytic energies and prior-floored component weights. Bottom right (BEHAVE): plan/event/contact features score the same fixed 16 trajectories. Dashed loss branches denote training supervision; the dashed plan-geometry route is an inference input. Architecture and information access are specified in Section A.2.

A.1 TASK, DATA, AND INFORMATION ACCESS

Rigid poses belong to the special Euclidean group $\mathrm{SE}(3)$; rotations belong to the special orthogonal group $\mathrm{SO}(3)$. The input is observed history and a supplied human plan; outputs retain each corpus's native object-state representation. Table 7 summarizes the notation used throughout the method; t, z, and k index future time, chart, and candidate, respectively.

TRUMANS uses 13 observed and 32 planned frames with a 44D object-relative human feature. Thirteen action pairs support training and seven support development, including loss and gate selection. Both branches share history and camera motion. The touch branch retains recorded motion; local inverse kinematics changes future right-arm joints 8–10 to create a 3–6 cm near miss. Its object target remains analytically at the boundary pose under the static rigid-object assumption. History agrees within 1 mm and 1 degree, and touch distance is below 2 cm.

BEHAVE supplies human plans in the Skinned Multi-Person Linear model with hands (SMPL-H) and native object poses: 9,952 training windows and 3,887 official-test windows from 82 sequences, with 13/32 observed/future frames. HUMOTO provides 5,775 training and 871 test windows from 11 test sequences; its primary score combines translation and point error. CORE4D uses 960 training and 217 confirmation windows from 24 confirmation sequences. HOI-M3 has 11,464 training windows and 1,579 final windows from three independent sequences. MMHOI uses four observed and four future frames; 3,369 training windows precede 563-window validation and 585-window test roles. Interval estimation groups windows by parent sequence or capture.

Plan and target provenance. MMHOI's object context comes from observed history and the boundary object center. Planned human centers are derived from actor bounding boxes, not object

Table 7: **Main method notation.** Inputs, intermediate responses, and candidate scores.

Symbol	Meaning
$h,\ a_{1:T},\ T$	Observed history, supplied human-motion plan, and prediction horizon.
$s_t,\ \bar{s}_t,\ f_\theta$	Target object state, predicted response, and response predictor.
(p_t, R_t)	Translation and rotation components of a rigid object state.
$b \in \{+,-\},\ \Delta\bar{s}_t$	Touch/near-miss branch and difference between predicted branch responses.
$q_t,\ \eta_t,\ E_t$	Event probability and cumulative event weights in the paired and full-body paths.
$z \in \mathcal{Z},\ d_{t,z}$	Coordinate chart and within-chart prediction dispersion.
$\Delta h_t,\ \Delta R_t^h$	Translation and rotation implied by the available hand-plan prefix.
$c_{t,z}^u,\ e_{t,z}^u,\ \pi_{t,z}^u$	Plan mismatch, energy, and fusion weight for $u \in \{p, R\}$.
$u_t^j,\ v_t^j,\ \alpha_t^j$	Learned response, contact transport, and temporal trust for $j \in \{p, r\}$.
$K,\ S_K(h, a)$	Candidate count and complete response trajectories for the given input.
(s_p, s_R, o_p, o_R)	Candidate scales and trust offsets for translation and rotation.
$x,\ r_\psi(x),\ p,\ w$	Plan evidence, preference logits, training probabilities, and inference weights.
$J_k,\ R_k,\ k^*$	Candidate joint error, normalized risk, and target-selected best index (training only).

target positions. Planned actor pose and annotated action edges form a supplied semantic plan; actor removal therefore measures attribution conditional on this plan. Replacing object targets leaves cached features and forward predictions unchanged in the raw-input audit. SSv2 uses offline, full-sequence wrist/object tracking and stabilization to obtain 2D trajectories. Of 2,016 source clips, 13 lack structural caches and 971 fail quality checks, leaving 745/143/144 training/development/test clips. The generated-plan cohort retains 69 of 120 clips after two missing-cache and 49 quality exclusions. Shared-video pairs form 104 connected resampling groups in the recorded-plan test. The reconstruction frontend and the target-free dynamics forward have distinct information boundaries.

A.2 Architecture and Learning

To distinguish interaction onset from continuous motion, the paired predictor uses two four-head prefix-Transformer blocks (width 32, feed-forward width 48) for geometric-touch evidence, and a width-48 GRU response path with two input layers and a relative 6DoF head. It contains 40,238 parameters. Paired branches share weights; all state, velocity, terminal, and difference losses supervise recurrent responses after cumulative sigmoid weighting and before external feasibility. Each seed first multiplies its state-head output by its own soft weight before averaging the five physical-unit responses. A separate attention ensemble supplies mask confidence (Section B.2).

BEHAVE response branches each contain 116,423 parameters. Eighteen landmarks contribute object-frame position, velocity, template displacement, and surface distance to a 180D input. Its auxiliary event denotes object response beyond 2 cm translation or 5 degrees rotation, whereas TRUMANS contact supervision denotes geometric touch. HOT3D chart experts contain 177,929 parameters and train for 4,000 updates per fold. The shared response reads history, plan prefixes, and expert response trajectories without identity or device labels.

The BEHAVE ranker maps 173D plan/event/contact summaries through a 64-unit sigmoid linear unit (SiLU) layer to 16 logits. Its loss combines normalized expected risk, oracle-codeword cross entropy (weight 0.1), and entropy regularization (weight 0.01). Per-example risks are divided by their mean, floored at 10^{-6}; the development-selected inference logit multiplier is 0.25. These candidate probabilities are distinct from contact probabilities and temporal trust. Specialized paired-predictor comparisons use five seeds and 2,500 updates with the same backbone and state objectives. Sequence-disjoint development selects checkpoints and ranking settings; results distinguish scores of ensemble predictions from metrics averaged over independently evaluated training seeds.

Ranking features and risk. The 173D vector concatenates 22 global transport/event descriptors, four summaries of 13 temporal channels (mean, sample standard deviation, maximum, and final value), and three summaries of 33 physical-context channels (mean, sample standard deviation, and

final value). Normalized time is excluded from the temporal summaries. The physical channels contain 22 observed object-motion, orientation, and shape descriptors, human-relative position and displacement (three each), and five body-motion or distance magnitudes. Object geometry is fixed at the observed boundary; plan-derived features are recomputed for every supplied plan. Each component undergoes $x \mapsto \operatorname{sign}(x)\log(1+|x|)$, followed by the saved fitting-set mean and standard deviation. The released feature index specifies the concatenation order and each operation.

For sample i and candidate k, let $E_{ik}^{p,A}, E_{ik}^{p,F}$ denote translation ADE/FDE in centimeters, and $E_{ik}^{R,A}, E_{ik}^{R,F}$ rotation ADE/FDE in degrees. For scored role $\mathcal{D}$, define $b_{\mathcal{D}}^{u,v}$ as the corresponding mean error of its deterministic reference (floored at 10^{-6}). The reported ranking cost is

$$J_{ik} = \frac{1}{2}\sum_{u\in\{p,R\}}\sum_{v\in\{A,F\}}\frac{E_{ik}^{u,v}}{b_{\mathcal{D}}^{u,v}}, \qquad R_{ik} = \frac{J_{ik}}{\max(K^{-1}\sum_j J_{ij}, 10^{-6})}. \tag{10}$$

Training uses R_{ik}; evaluation reports $\sum_k w_{ik}J_{ik}$ and the same J for argmax and oracle comparisons. On the official role the four reference scales are 6.5693 cm, 11.1374 cm, 13.5463°, and 22.6444°. These are role-level scoring statistics, not ranker inputs; development fitting uses its own reference errors. Joint-ADE and joint-FDE coverage use the corresponding two-term sums separately. Thus their oracle choices can differ from the four-component ranking oracle. Ranking uses tensor-based geodesic errors; state and generated-plan tables use SciPy rotations. For the same deterministic prediction, their ADE/FDE differ by only $0.00034/0.00031^\circ$ between implementations.

A.3 Detailed Structured Inference

Temporal trust and persistent transport. For BEHAVE, let $x_{t,l}$ be planned landmark l in the observed boundary object frame and $d_{t,l}$ its nearest-template-vertex distance. Geometry stays referenced to the observed object pose, rather than following either a predicted or target object. The 18 landmarks receive weights proportional to $\exp(-\min_t d_{t,l}/0.1)$; only the largest eight remain, renormalized to sum to one. Weighted Kabsch maps the first planned landmark set $x_{1,l}$ to each $x_{t,l}$, with a $10^{-6}I$ covariance regularizer and determinant correction. Its local translation is rotated by the boundary object rotation into world-axis displacement; its rotation is converted to an axis-angle vector. Together, these components define the transport v_t used in response composition.

Learned branches supply displacement u_t^p and relative axis-angle u_t^r. The trust network receives 22 complete-plan summary features and 14 time-resolved features: rigid-fit residual, contact distance, anchor motion, transport magnitude, response–transport disagreement, event probability and its cumulative maximum, and normalized time. Nonnegative features undergo $\log(1+x)$ and fit-role standardization. A 32-unit SiLU layer with two sigmoid outputs predicts α_t^p, α_t^r; the reported system averages its seed-specific trust fields. Both weights range over $[0,1]$. Equation 8 combines these fields with cumulative response-event evidence E_t. The final pose is $p_t^o = p_0^o + \bar{s}_t^p$ and $R_t^o = R_0^o\exp([\bar{s}_t^r]_\times)$. The fused axis-angle coordinates are exponentiated after linear combination of the two sources. The fixed-transport control uses constant component weights, and contact geometry uses the learned response without this adaptive transport mixture. Anchors remain fixed across the horizon; release can change the learned response and trust, but there is no explicit contact-release state machine or anchor reinitialization. The minimum-distance anchors and global summaries use the complete supplied plan, with geometry fixed at the observed boundary.

Chart mappings. Write O_0 for the boundary object transform and W_0^H, W_t^H for boundary and planned wrist transforms of hand H. The contact chart represents hand history and plan by $O_0^{-1}W_t^H$; the wrist chart uses $(W_0^H)^{-1}W_t^H$. The factorized chart combines contact-frame hand history with wrist-anchored planned motion. Each hand contributes translation, the first two rotation columns, translation norm, and visibility. The common object-history stream contains $O_0^{-1}O_t$ and boundary-wrist-relative hand history. All three experts directly predict the same target $O_0^{-1}O_t$ (translation in meters and 6D rotation columns), rather than predicting in their respective input frames and transforming outputs afterward. World reconstruction left-multiplies the predicted relative transform by O_0. Observed O_0, W_0^H anchor reconstruction; planned W_t^H supplies motion.

For hand $H \in \{L, R\}$, object-relative poses use $(\boldsymbol{T}_0^o)^{-1}\boldsymbol{T}_t^{H,b}$ and the camera uses $(\boldsymbol{T}_0^C)^{-1}\boldsymbol{T}_t^C$. The 44D feature concatenates bilateral translation, velocity, 6D rotation, relative camera motion, hand–object geometry, object context, and time; the 180D full-body form uses the 18 landmarks described

above. For chart $z \in \{c, w, f\}$, seed mean $\bar{s}_t^z$, boundary distance δ_k, visibility v_k, and planned hand motion, the complete compatibility terms are

$$\begin{aligned} \omega_k &\propto e^{-\delta_k/\tau_h}(0.1 + 0.9v_k), & \textstyle\sum_k \omega_k &= 1, \\ \Delta h_t &= \textstyle\sum_k \omega_k (p_t^k - p_0^k), & c_{t,z}^p &= \mathrm{Z}_{\mathcal{Z}}(\|\bar{p}_t^z - \Delta h_t\|_2), \\ c_{t,z}^R &= \mathrm{Z}_{\mathcal{Z}}(d_{\mathrm{geo}}(\bar{R}_t^z, \Delta R_t^h)), & \kappa_t &= (1 - e^{-d_{\mathrm{geo}}(I, \Delta R_t^h)/\tau_R}) e^{-H(\omega) d_{\mathrm{geo}}(Q_t^L, Q_t^R)/\tau_a}, \\ e_{t,z}^p &= \mathrm{Z}_{\mathcal{Z}}(d_{t,z}) + \alpha g_t c_{t,z}^p, & e_{t,z}^R &= \mathrm{Z}_{\mathcal{Z}}(d_{t,z}) + \alpha g_t (c_{t,z}^p + \beta \kappa_t c_{t,z}^R), \end{aligned} \tag{11}$$

where $k \in \{L, R\}$ indexes hands, p_t^k, R_t^k are planned poses, δ_k and v_k are boundary hand–object distance and visibility, and $\delta_{t,k}$ is planned distance. $Q_t^k = R_t^k (R_0^k)^\top$, I is the identity rotation, ΔR_t^h is their ω-weighted chordal mean, and $g_t = e^{-\min_k \delta_{t,k}/\tau_g}$. The lower-left panel of Figure 5 abbreviates $\mathrm{Z}_{\mathcal{Z}}(d)$ as $\widetilde{d}$ and suppresses the fixed α, β multipliers in its energy tiles. "Consensus" κ_t combines rotation magnitude and bilateral agreement; the proximity term g_t controls the overall plan contribution. Standardization subtracts the across-chart mean at fixed sample and time and divides by the across-chart root-mean-square (RMS) deviation, floored at 10^{-6}. "Prior-floored Softmax" denotes $\pi^u = 0.4\pi^0 + 0.6\,\mathrm{softmax}(\log \pi^0 - e^u)$, applied to each component. These analytic weights add no parameter, optimizer update, or expert forward. The frozen settings are $\alpha = 0.75$, $\beta = 3$, $\tau_h = 0.03$ m, $\tau_g = 0.24$ m, $\tau_R = 10°$, $\tau_a = 30°$, and $\pi^0 = (0.40, 0.25, 0.35)$ in contact/wrist/factorized order. Hand entropy is $H(\omega) = -\sum_k \omega_k \log \omega_k / \log 2$. With M ensemble members indexed by m, the implemented dispersion is

$$d_{t,z} = \sqrt{\max\left(\frac{1}{M}\sum_m \|p_t^{z,m} - \bar{p}_t^z\|_2^2, 10^{-12}\right)} + \frac{0.1}{M}\sum_m d_{\mathrm{geo}}(R_t^{z,m}, \bar{R}_t^z). \tag{12}$$

Equation 12 uses member m's translation $p_t^{z,m}$ and rotation $R_t^{z,m}$. Translation is in meters and angular distance in degrees; the fixed coefficient 0.1 m/degree maps angular dispersion to a translation-equivalent score before across-chart standardization. This coefficient fixes the relative contribution of the two terms before standardization. Bars denote ensemble means; rotation columns are averaged across members and then orthonormalized using ρ_6 below.

For rotation, each chart predicts the first two columns $u_z \in \mathbb{R}^6$. We average these columns in $\mathbb{R}^6$ and reconstruct a valid rotation matrix using Gram–Schmidt orthonormalization:

$$\rho_6(u) = \big[r_1,\, r_2,\, r_1 \times r_2\big], \qquad r_1 = \frac{u_1}{\|u_1\|}, \qquad r_2 = \frac{u_2 - (r_1^\top u_2) r_1}{\|u_2 - (r_1^\top u_2) r_1\|}. \tag{13}$$

In Equation 13, $u = [u_1; u_2]$ stacks two three-dimensional columns; r_1, r_2 are orthonormalized columns. Weighted plan-implied transport uses determinant-corrected singular value decomposition (SVD) projection; chart-posterior fusion uses ρ_6. Figure 2(b) groups two response heads as shared g_ϕ: each reads history and plan prefixes, with translation reading the chart mean and rotation reading expert columns and chart weights. The 6D residual concatenates translation and a rotation vector, with three coordinates per component. Raw bounds are 6 cm and 12° (converted to radians). Translation adds directly; rotation composes by right multiplication in $\mathrm{SO}(3)$.

Population evaluation contracts translation by $\lambda = 0.5$. Rotation instead averages the two-column outputs of a learned router and a bounded-response model before applying ρ_6; both use a 0.25 prior floor, and only the latter predicts a nonzero rotation residual. The figure's common λ scaling summarizes attenuation: translation contracts its residual, whereas rotation blends the two models' output columns before orthonormalization. Incoming weights include frozen one-step rotational-tangent compatibility; routing and component heads retain separate checkpoints. The analytic chart-only comparison uses frozen experts without these learned heads.

Multi-actor composition. In MMHOI, shared actor encoders receive relative centers, center increments, pose, semantic action embeddings, and interaction masks, with time but no actor-index embeddings. Future queries attend to valid actor and observed-object tokens. Each actor proposes its planned center displacement plus a bounded learned correction. Three global alternatives supply a learned displacement, zero displacement, and constant velocity. A softmax over all actor and global alternatives defines their weighted sum. Actor logits combine query–actor similarity with

the supplied action-edge prior; invalid actors receive a -10^4 logit. Consequently, permuting poses, semantic features, and masks together permutes attribution weights while preserving the response. On a separate 763-window external-validation cache, three permutations across three seeds change predicted translation by at most 6.6×10^{-7} m. This conditional attribution uses the supplied semantic edge. Complete-wrist and one-hand release/join transport variants instead use weighted Kabsch transport and then predict a residual correction from the supplied input.

A.4 Inference and Reproducibility

Inference constructs observed-geometry/plan features before response, transport/chart, and ranking. Changing a plan recomputes distances, anchors, transport, event, and trust; local prefix masks do not restrict complete-plan statistics or ranking. Table 8 includes normalization, host-to-device transfer, prediction, and host output on a shared NVIDIA L20, starting from structured features. Three repetitions on the first sixteen serialized windows interleave methods after warmup. Surface extraction, RGB reconstruction, disk access, and rendering are outside this timing boundary.

Table 8: **Structured prediction cost.** WorldEffect issues 16 futures in 37.7 ms median. Batch one; 48 interleaved runs per method; timings are in milliseconds.

System	Parameters	K	Neural calls	Median	90th pct.
Transformer ensemble	262,485	1	3	5.0	6.6
Conditional flow	288,678	16	$1 + 32$	29.1	36.5
Conditional diffusion	288,678	16	$1 + 32$	62.0	83.1
WorldEffect	1,001,301	16	15	37.7	50.0

Flow/diffusion use one encoder and 32 field/denoiser calls processing all sixteen candidates. Their timing is per seed, whereas accuracy averages three separately scored distributions. Transformer and WorldEffect timings include their deployed three-seed ensembles and all branches.

B Additional Experimental Results

B.1 Metrics and Statistical Units

Translation ADE/FDE use centimeters; rotation uses geodesic degrees. Each oracle selects one complete trajectory, never a per-frame mixture. Component-wise minima may use different translation and rotation candidates; coherent joint minima use one candidate and frozen deterministic-error normalization. Relative error reduction divides by the baseline error. Aligned reduction divides by reversed error, whereas reversal penalty divides by aligned error.

The translation energy score is $\mathrm{ES} = \mathbb{E}\|X - y\| - \frac{1}{2}\mathbb{E}\|X - X'\|$, where complete trajectories in centimeters are vectorized and divided by $\sqrt{T}$. The finite issued distribution includes diagonal pairs. Each flow seed issues 16 samples and is scored before seed averaging. Brier score sums squared class-probability errors; negative log-likelihood (NLL) uses the oracle codeword label; expected calibration error (ECE) uses ten fixed confidence bins. These assess different distribution properties. Projection metrics use calibrated vertices, pixel trajectory/endpoint errors, target-box-normalized projection error (NPE), and hull IoU. Bootstrap draws preserve paired independent sequences or captures and the stated aggregation, including within-unit windows for window-micro estimates. The component comparisons fix the reported ensembles: without the external mask, paired-loss Pair ADE improves by 21.93% (95% CI $[7.79, 42.78]$), temporal-trust T-ADE by 16.01% ($[12.93, 19.47]$), and same-$K = 16$ trust–response versus magnitude-cover T-minADE by 11.75% ($[10.49, 13.06]$). BEHAVE intervals use 5,000 paired sequence draws and recompute relative reduction from each resample's mean errors; they exclude training-seed uncertainty. The paired-loss interval enumerates 7^7 action-pair resamples, reflecting cohort variation.

B.2 Paired Effects and Feasibility

Pair-conditioned feasibility. To suppress unsupported motion in the controlled-pair evaluation, let d_t^b be the smaller distance of the two planned wrists to the observed boundary-object origin,

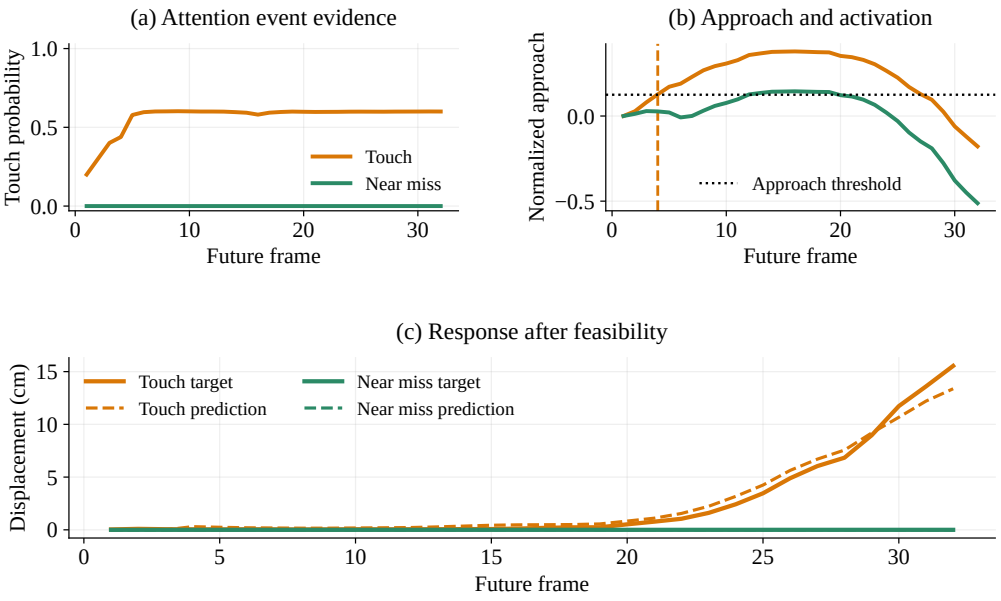


Figure 6: **Contact-contrastive feasibility.** (a) Attention-event probabilities supply branch confidence. (b) Normalized approach, its threshold (dotted), and mask onset (dashed). (c) Measured/constructed targets (solid) and gated predictions (dashed). The previously displayed development case is retained; its original selection uses terminal effect error and direction.

in meters, and define $g_t^b = (\max(d_1^b, 0.05) - d_t^b)/\max(d_1^b, 0.05)$. The attention event ensemble supplies $\tilde{q}_t^b$; its complete-plan mean is $c_b = T^{-1}\sum_t \tilde{q}_t^b$. With $b^* = \arg\max_b c_b$, the frozen rule is

$$E_t^b = \mathbf{1}\left[d_t^b \le 0.10 \ \vee \ (g_t^b \ge 0.125 \ \wedge \ b = b^* \ \wedge \ |c_+ - c_-| > 0.05)\right], \qquad M_t^b = \max_{\tau \le t} E_\tau^b. \quad (14)$$

Proximity or confidence-qualified approach activates the remaining horizon; equal confidence gives no approach activation. Thresholds use consumed development roles. The mask follows ensemble averaging and retains activation; BEHAVE separately models release-dependent transport.

Table 9 varies paired learning and the external mask while retaining soft weighting. Reweighting doubles nonzero-translation-frame weights, normalized to mean one. Without the mask, paired learning reduces touch ADE (3.410 to 2.863 cm) and magnitude error (10.751 to 9.709 cm), while direction cosine decreases. Cosine uses a 10^{-12} denominator floor; zero predicted effects contribute zero. Leave-one-pair ADE gains span 13.68–29.70%, compared with 21.93% on all seven pairs, quantifying sensitivity to individual development actions.

Table 9: **Learning and feasibility.** Paired learning and gating both reduce Pair ADE on seven development pairs. Matched five-seed backbones and 2,500 updates retain soft weighting; Gate toggles the external mask after ensemble averaging.

Objective	Gate	Pair ADE↓	Touch FDE↓	Near ADE↓	Effect error↓	Cos.↑
State	Off	1.939	11.909	0.468	12.622	**0.854**
State	On	1.641	11.909	0.000	11.909	0.739
Paired	Off	1.514	**11.516**	0.164	11.774	0.810
Paired	On	**1.407**	**11.516**	**0.000**	**11.516**	0.742
Reweighted	Off	1.918	12.079	0.656	13.013	0.833
Reweighted	On	1.531	12.079	0.000	12.079	0.750

Figure 6 illustrates this distinction: cumulative feasibility suppresses the constructed near-miss response while preserving motion on the supported touch branch.

Composition of the three mechanisms. We train paired chart responses, fuse them, and rank a cover generated from that fusion (Table 5). Thirteen training pairs from three scenes supply supervision; scene-held-out training predictions fit the cover and ranker. Three final seeds use 2,500 updates each. The seven development and fourteen diagnostic pairs each span three other scenes; both are previously examined constructed-pair cohorts. The recurrent response uses the cumulative soft weight of Equation 2; fusion and ranking operate without the external binary mask.

One width-48 GRU shares paired branches and three chart-coded views: object frame, observed boundary-right-wrist frame, and wrist-frame positions with object-frame orientations. All outputs use the boundary-object frame. Component weights are $0.9\,\mathrm{softmax}(-\mathrm{Z}_{\mathcal{Z}}(c)) + 0.1/3$, where c compares the response with event-weighted planned wrist transport. Translation is averaged; determinant-corrected SVD projects the weighted rotation matrices to $\mathrm{SO}(3)$.

Table 5 uses the final score-trained pipeline, with frozen downstream codewords and router for all removals. No-pair experts use matched-budget retraining; equal-chart weights also change the fused-response ranker features. Joint risk normalizes translation and rotation ADE by training zero-motion means, 2.3198 cm and $3.8097°$. The full system in Table 1 refines ordering with direct distribution scoring. Translation/axis-angle scales $\{0, 0.25, 0.5, 1, 1.5\}$ and event-logit offsets $\{-1, 0, 1\}$ form 225 codewords; offsets precede sigmoid and cumulative maximum. Greedy joint-ADE coverage retains identity, zero response, and fourteen further codewords. A linear softmax (W, b) uses 24 descriptors: mean/min/max bilateral and minimum wrist distances (9), right-wrist speed mean/max/final (3), and fused state mean/final (12). Its objective is $\mathrm{ES}/s_{\mathrm{ES}} + 0.25\sum_k w_k R_k/2 + 0.01[\mathrm{mean}(W^2) + \mathrm{mean}(b^2)]$, where R_k is joint ADE above and s_{ES} is training zero-response ES. Nested training-scene selection compares risk weights $\{0.25, 1\}$ and regularization $\{0.01, 0.1\}$ using normalized T/R ADE/FDE and ES. The selected ranker uses 600 updates; inner/outer scene exclusion requires eighteen additional response fits.

For a single output, the integrated system selects $\hat{k} = \arg\min_k \sum_j w_j[\mathrm{ADE}_p(S_k, S_j)/2.3198 + \mathrm{ADE}_R(S_k, S_j)/3.8097]$, using the unrounded training scales. Selected on held training scenes, this target-free whole-trajectory rule is shared across Table 1 and leaves the distribution unchanged.

Matched-budget refitting of the cover and router after component removal gives diagnostic ADE of 2.273 cm with equal chart weights, matching the full model at the reported precision; the full model retains lower diagnostic weighted risk and ES. Development scores favor reduced variants, so the frozen-component comparison measures dependence on the released configuration.

Matched-input predictive systems. Table 1 gives all predictors the same three views and planned transport (138 channels). GRU/Transformer use paired supervision; the learned set uses best-of-set regression and classification. All average three seeds trained for 2,500 fixed updates; corresponding set candidates retain $K = 16$. Training-only anchors have fourteen distinct values because static targets repeat. WorldEffect uses nine held-scene fits to learn its cover and router.

A frozen-source check uses seven recorded windows from three further TRUMANS recordings and three constructed pairs from one of them. WorldEffect leads recorded ADE and weighted risk, while the learned set has lower ES (10.440 versus 10.507 cm). On the constructed pairs, the learned set leads ADE, risk, and ES: 1.435/1.430/2.255 versus WorldEffect's 2.195/1.717/3.170. All predictors retain their original checkpoints; recording-level aggregation gives equal source weights.

Response differences in recorded interactions. We match recordings within object category across sequences, with planned human-root paths at least 5 cm apart (RMS). Thirteen-frame object-local root, object displacement, pose, and gravity histories use dimension-normalized scales of 0.25 m, 0.05 m, 0.5 rad, and 0.25. A training nearest-neighbor quartile sets the threshold; greedy matching uses each sequence once without object futures. The 22 pairs cover 44 previously evaluated sequences and three participants; mean history mismatch is 5.92 cm in root position, 0.63 cm in object position, and 0.127 rad in non-root pose parameters over the complete observed history.

Table 10 compares response differences in boundary-object coordinates. Cosine is zero when either terminal vector is at most 1 cm; all pairs remain, including six near-zero measured differences. WorldEffect improves difference accuracy over the plan Transformer. This observational comparison retains initial-state mismatch rather than assuming identical histories.

Table 10: **Measured response differences.** WorldEffect improves all three metrics on 22 outcome-blind matched recording pairs. Errors: pair means in cm; cosine: terminal direction agreement.

Method	Difference ADE↓	Terminal error↓	Cosine↑
Zero motion	8.18	19.37	0.000
Constant velocity	7.67	18.74	0.330
Plan Transformer	5.90	13.59	0.403
WorldEffect	5.63	12.46	0.496

Executed alternatives from identical simulated states. We use MuJoCo (Todorov et al., 2012) to execute both branches after restoring the same settled integration state. A prescribed spherical hand proxy (radius 1.5 cm) interacts with a $10 \times 8 \times 5$ cm rigid cuboid on a plane. Paired plans contrast pushing with bypassing, stopping before contact, or pushing from the opposite side. Both object trajectories come from physics integration, including the noncontact branch. This isolates contact-induced responses in a synthetic wrist-proxy setting with precisely controlled initial conditions.

Disjoint random seeds generate 192/48/96 training/development/confirmation configurations, varying unobserved mass (0.08–0.35 kg), friction (0.25–0.80), and contact offset (± 2.5 cm). All roles contain the three plan families, with 13 static history and 60 future frames at 10 Hz. Matched width-48 GRUs use three seeds, 2,500 fixed updates, soft event weighting, simulated-contact labels, and no external mask. State/event/velocity/endpoint losses and initialization are shared; the paired loss is toggled. History-only uses boundary poses in place of planned hand poses.

Table 11 separates trajectory and response-difference errors. Cosine is zero if either terminal difference is at most 1 mm, retaining every pair. Sensitivity S divides the mean predicted terminal difference norm by its measured counterpart, with a 1-mm denominator floor. Plan-conditioned models recover direction and magnitude; the six-second paired-loss improvement over state supervision is 0.003 cm (paired configuration-bootstrap 95% CI $[-0.028, 0.028]$ cm). At two seconds it is 0.030 cm ($[0.014, 0.046]$ cm). These intervals resample initial configurations after averaging seed scores. Across the three action families (32 configurations each), plan conditioning consistently helps, while paired-loss gains vary. Opposite pushes have cosine 0.99997 but terminal magnitude error 1.648 cm, separating direction agreement from accurate response amplitude.

Table 11: **Executed action alternatives.** Plan-conditioned models recover responses from 96 held-out states. Three-seed means at six seconds; errors: cm; $S = 1$ matches effect magnitude.

Objective/input	State ADE	Difference ADE	Difference FDE	Cosine	S
History only	7.277	13.305	16.538	0.000	0.000
State objective + plan	0.441	0.842	1.024	0.997	0.999
Paired objective + plan	0.440	0.839	1.045	0.997	0.996

Objective weighting. A matched three-seed follow-up doubles state-loss weights above 1-mm training displacement, normalized by their training mean. Six-second state/difference ADE is 0.436/0.832 cm on the same evaluated configurations; the difference-ADE gap to paired learning is -0.007 cm (95% CI $[-0.035, 0.019]$ cm).

Exact replay reproduces targets; halving the integration step on six training configurations changes positions by 0.174 cm on average and 0.705 cm at most over both executed branches.

B.3 Full-Body Prediction, Coverage, and Ordering

Table 12 retains the learned and structured candidate comparisons omitted from the compact main table. Each oracle chooses a complete trajectory separately for each metric; all rows use the same 3,887 official test windows and sixteen candidates.

Issued trajectories and coverage. On the same 3,887 test windows, expected-utility argmax has joint risk 2.100 versus the single-trajectory reference's 2.000, a 5.00% increase (95% sequence-bootstrap interval $[3.13, 6.87]\%$). These single-output errors are distinct from probability-weighted

Table 12: **Matched candidate-count comparison.** The trust–response cover improves all BEHAVE oracle errors at $K = 16$. Translation: cm; rotation: degrees.

Method	T-minADE	T-minFDE	R-minADE	R-minFDE
Learned trajectory set	8.084	11.186	12.632	17.436
Conditional diffusion	7.283	10.867	14.262	20.352
Conditional flow	6.591	9.779	12.877	18.205
Same-base residual flow	5.570	7.905	11.757	15.995
Magnitude cover	5.250	8.115	11.376	17.718
Trust–response cover	4.633	6.702	10.318	15.120

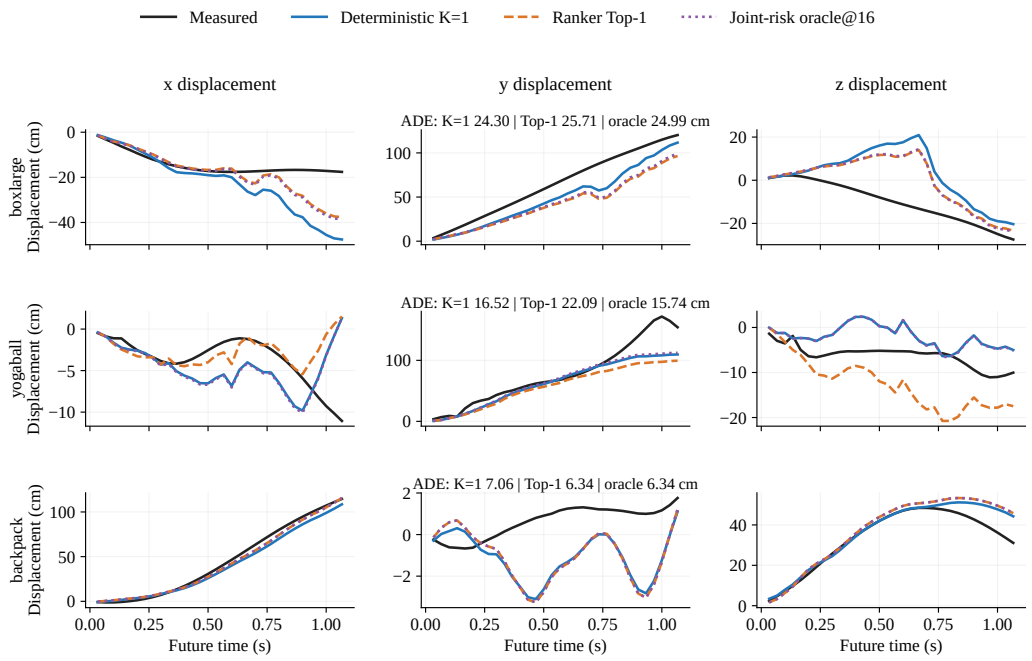


Figure 7: **Issued predictions and candidate coverage on identical cases.** Black is measured motion, blue is the deterministic response, orange is the saved ranker argmax, and purple is the joint-risk oracle from the same 16 candidates. Each candidate is fixed for the complete trajectory. Rows follow the first three previously fixed gallery cases, without selection on prediction error. The row annotations give translation ADE; the three columns expose all displacement components.

risk. Figure 7 shows these outputs on the first three cases of the existing target-only gallery protocol. Coordinates share the time axis and measured reference; oracle selection minimizes joint 3D risk.

The fixed cover contains 16 quadruples (s_p, s_R, o_p, o_R). Translation and rotation scales range over $[0.25, 1.65]$ and $[0.25, 1.75]$, respectively; trust offsets range over $[-0.5, 0.5]$. The released candidate artifact retains all 64 constants; no candidate is adjusted at test time. For codeword k and component $j \in \{p, r\}$, the actual trajectory generator is

$$\alpha_t^{j,k} = \mathrm{clip}(\alpha_t^j + o_{j,k}, 0, 1), \qquad s_t^{j,k} = s_{j,k} E_t\big[(1 - \alpha_t^{j,k}) u_t^j + \alpha_t^{j,k} v_t^j\big], \tag{15}$$

where $s_{r,k} = s_{R,k}$ and $o_{r,k} = o_{R,k}$. The same constants apply at every timestep; rotation reconstruction uses $R_0^o \exp([s_t^{r,k}]_\times)$. Only parameters and the generation rule are fixed across inputs, not the resulting absolute trajectories. Development grid searches identify separate translation/rotation scale–offset optima for ADE and FDE. Their unique quadruples, together with 16 magnitude-only cluster centers with zero offsets, form the candidate pool. Starting from an empty set, greedy facility selection alternates the mean best joint ADE and mean best joint FDE (ADE first), adding the candidate that minimizes the current objective. Sequence-disjoint development folds select this

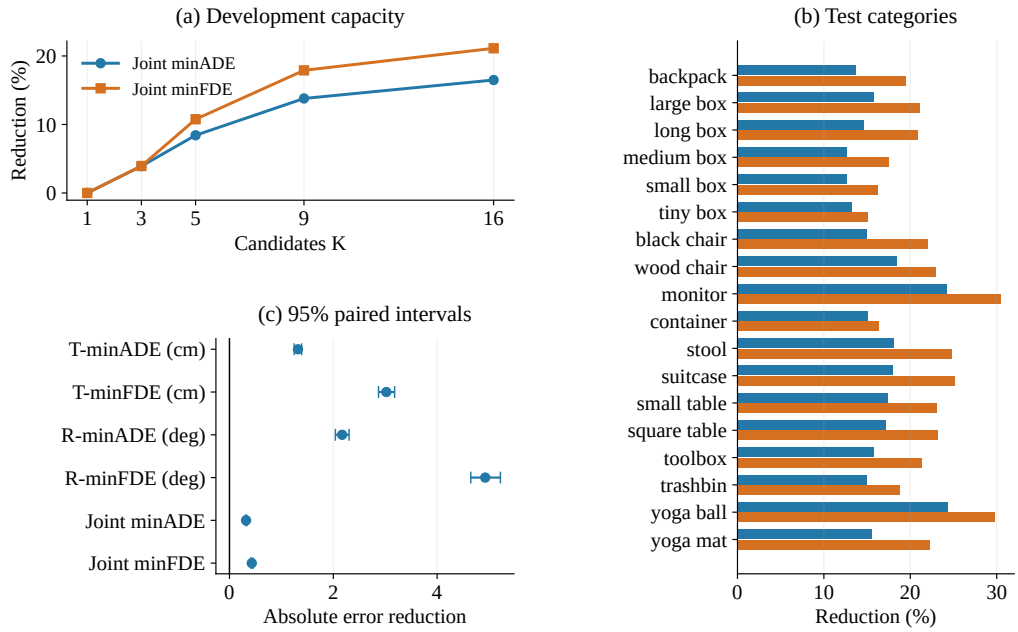


Figure 8: **Fixed-cover capacity.** Independently fitted covers at $K \in \{1, 3, 5, 9, 16\}$ reduce coherent joint minADE/minFDE. Blue/orange denote ADE/FDE in (a,b). All 18 test categories benefit at $K = 16$; sequence-bootstrap intervals remain positive. Development fixes construction and K.

construction; the final pool uses all 2,293 development windows. Here VQ quantizes response parameters through fixed selection rather than an encoder trained to assign codewords.

Figure 8 examines candidate count and per-category coverage. Each count uses its registered construction; non-nested sets need not improve monotonically.

Calibrated projection gives fixed transport 33.42 px trajectory error, 58.35 px endpoint error, and 0.5908 IoU. Temporal trust gives 28.48 px, 48.19 px, and 0.6544. The screen-space oracle cover gives 20.84 px, 35.02 px, and 0.7211. Candidate selection minimizes pixel rather than 3D error.

Learned uncertainty controls. With matched inputs, sixteen samples, and a 16-step solver, increasing flow capacity from 288,678 to 1,038,150 parameters changes T-minADE from 6.591 to 6.740 cm and ES from 8.204 to 8.289 cm. Three same-base residual heads fit two-fold sequence-disjoint base predictions; cover lowers ES by 0.655 cm (82-sequence CI $[0.493, 0.838]$ cm).

Conditional diffusion and sampling cost. Conditional diffusion (Ho et al., 2020; Song et al., 2021) uses the same inputs and 288,678-parameter backbone as flow: 1,000-step cosine noise training and 32-step deterministic denoising diffusion implicit model (DDIM) sampling. Sequence-disjoint development ES selects 6,000 of 1,500/3,000/6,000 updates. Three refits each issue $K = 16$; seed scores are averaged. Test ES is 9.327 cm versus uniform cover's 6.033 cm (difference 3.294 cm, paired 82-sequence 95% CI $[2.740, 3.874]$ cm).

Input-dependent probability ordering. Table 13 fixes candidates and varies probabilities. Ridge regresses candidate errors from standardized features; negative-error softmax gives probabilities. Regularization and temperature use development data. The global prior fits sixteen input-independent logits on 2,293 development windows with the same utility objective. Five within-category permutations either mismatch all ranker features or only 74 plan-geometry channels, retaining each recipient's candidates and targets. The latter retains context, event predictions, learned response, and response-disagreement channels from the original recipient sample.

Expected utility improves ES over the global prior by 0.124 cm (CI $[0.079, 0.166]$) and plan-geometry mismatch by 0.108 cm ($[0.080, 0.136]$), resampling 82 sequences with window-micro aggregation. The 12.73% reduction concerns joint risk; ECE rises despite better Brier/NLL. Post-hoc

Table 13: **Distribution scoring.** Learned weighting lowers joint risk but raises ECE. The same 16 candidates cover 3,887 BEHAVE windows; ES: translation only (cm); lower is better.

Ranking	Joint risk	ES	Brier	NLL	ECE
Uniform	2.688	6.033	0.9375	2.773	0.0126
Global utility prior	2.581	6.042	0.9387	2.770	0.0494
Ridge	2.437	6.005	0.9262	2.693	0.0165
Mismatched utility inputs	2.555	6.183	0.9438	2.931	0.0379
Mismatched plan geometry	2.431	6.025	0.9145	2.642	0.0119
Expected utility	**2.346**	5.917	**0.8937**	**2.479**	0.0581
+ NLL temperature	2.460	**5.898**	0.9071	2.550	0.1091

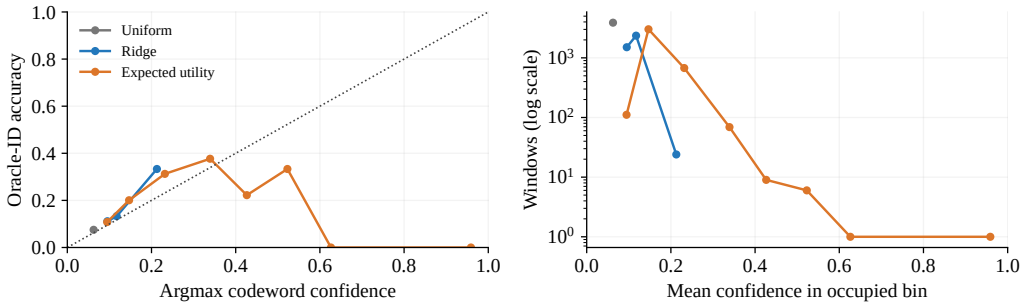


Figure 9: **Oracle-codeword reliability and confidence support.** Left: empirical argmax accuracy versus confidence on 3,887 windows; the dotted diagonal denotes perfect calibration. Right: the number of windows in each occupied bin. Both panels use the unchanged published probabilities and ten fixed bins. Empty bins are omitted, and connecting lines are visual guides.

NLL fitting on 2,293 development fold-held-out predictions selects inverse temperature 0.1364 from $[0, 4]$, versus published 0.25; cover/ranker settings already use development. On the previously evaluated official cohort, ECE, NLL, and risk worsen; ES changes by -0.019 cm (CI $[-0.047, 0.005]$). Argmax is unchanged, and published probabilities are retained. Figure 9 uses ten equal-width argmax-confidence bins and minimum-risk codeword agreement (first index for ties). Populated utility bins are mostly underconfident, with a sparse high-confidence tail. This discrete calibration does not measure physical safety; confidence-based rejection requires separate validation.

B.4 Generated and Imperfect Plans

Table 14 fixes the response model and recomputes features for every supplied plan. Ridge maps 13 poses to 32 via 64 training-fitted principal components (development-selected regularization 1,000). Root ADE improves from persistence's 13.13 to 9.78 cm, but object error increases. A subsequent relative-rotation forecaster improves object ADE to 14.845 cm on this already examined cohort. A 173,599-parameter GRU, selected at 4,000 updates and refitted with three seeds, lowers root ADE to 8.79 cm, but wrist ADE is 23.71 cm versus persistence's 23.55 cm. Its object improvement is 1.04% (CI $[-1.90, 3.50]\%$). Persistence is a static input plan, not a retrained history-only model.

Figure 10 uses wrist-ADE boundaries of 10, 20, and 40 cm; groups contain 376/1,584/1,765/162 windows. Forecast plans improve the first two groups by 0.82 and 1.63 cm (CIs $[0.24, 1.55]$ and $[0.96, 2.41]$). Larger-error groups are inconclusive. Recorded-plan error also changes across groups, linking this association to interaction difficulty as well as plan quality. Measured wrist-plan error defines these analysis groups, but is never an inference input.

Table 15 evaluates 33 previously unused HOT3D source captures: 749 Quest 3 and 932 Aria windows from four and seven identities. One of 34 deterministically sampled captures has no window meeting the established motion threshold. A ridge forecaster fits existing training windows, with regularization selected on development; all conditions share response experts and recompute plan

Table 14: **Generated-plan response.** Lower root error need not improve object response. The same checkpoint scores 3,887 BEHAVE windows; T/R: cm/degrees. Relative-rotation and GRU rows are follow-up diagnostics; the GRU averages three plan seeds.

Supplied plan	T-ADE↓	T-FDE↓	R-ADE↓	R-FDE↓
Recorded plan	6.569	11.137	13.547	22.645
History-persistent human pose	15.796	28.363	20.651	37.420
History-only motion forecaster	17.660	32.158	24.331	43.260
Relative-rotation forecaster	14.845	27.832	19.173	34.372
Nonlinear GRU forecaster	15.633	28.814	21.854	39.319

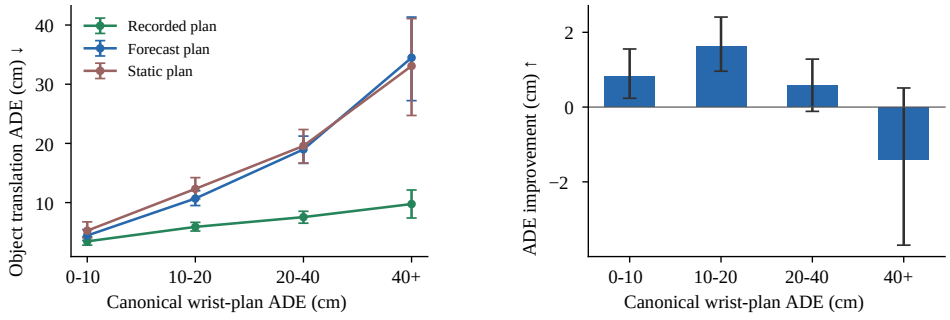


Figure 10: **Frozen 3D plan-quality analysis.** Left: response error within wrist-plan-error groups. Right: paired improvement over a static plan; positive values favor forecasting. Bars show 95% sequence-bootstrap intervals. All 3,887 windows are included.

geometry. Compatibility improves translation over fixed weights under the same forecast by 12.17% and 8.21%. Forecasting itself improves over static input on Aria but not Quest 3.

Table 15: **New-capture confirmation.** HOT3D translation/rotation ADE (↓) use cm/degrees, averaged equally over clips within identity and then over identities.

Plan / response	Quest T	Quest R	Aria T	Aria R
History constant velocity	3.49	16.60	6.01	24.84
Recorded / fixed prior	3.42	14.25	4.75	21.42
Recorded / compatibility	3.01	13.68	4.19	20.80
Static / fixed prior	3.81	14.47	5.82	22.55
Static / compatibility	3.35	14.18	5.42	22.26
Forecast / fixed prior	4.01	14.73	5.51	22.46
Forecast / compatibility	3.52	14.57	5.05	22.20

B.5 Video-Derived and Multi-Actor Interactions

The SSv2 quality protocol retains 1,032/2,016 processed videos (51.2%): 745/1,440 training, 143/288 development, and 144/288 evaluation. From 2,003 complete caches, wrist, object-track, rigid-fit, background, camera-validity, and camera-inlier gates retain 1,889/1,095/1,088/1,072/1,071/1,032 videos. Retention ranges from 21.7% for throw-success to 79.5% for edge-lift/drop. Object-track survival fails most often (846 videos; failures overlap). Accuracy is conditional on reliable reconstructed targets, and retained action proportions differ from the source pool.

For generated SSv2 plans, a five-seed GRU predicts 15 wrist states from seven observed frames and a supplied coarse action family. Reliability uses predicted contact, ensemble disagreement, dispersion, and observed context. The two replication roles exclude the original 2,016 clips and development data; their 69 accepted videos form 49 connected pair clusters. Pooled object ADE changes from 27.38 to 26.21 px (4.28% reduction); the cluster-resampled estimate is 4.87% with interval $[1.91, 7.76]\%$. These aggregation rules remain distinct. On the two roles (34/35 clips), history-

only ADE is 29.38/25.43 px, fixed composition 29.10/25.07 px, generated response 28.59/24.34 px, and reliability composition 28.34/24.13 px. All response variants reuse the same generated wrist files. Reliability reduces stationary false positives but lowers dynamic quality score (DQS) from 0.3973/0.4455 to 0.3358/0.3740. DQS is the geometric mean of trajectory-vector overlap and movement F1 score (the harmonic mean of precision and recall). Overlap combines positive direction cosine with the smaller-to-larger displacement-magnitude ratio over visible points and frames. Movement F1 uses a 3 px threshold and assigns zero to a zero denominator; DQS requires at least five moving target points. It complements pixel ADE.

Figure 11 illustrates generated plans and downstream trajectories. Its caption specifies performance-based selection, distinct from input-quality screening.

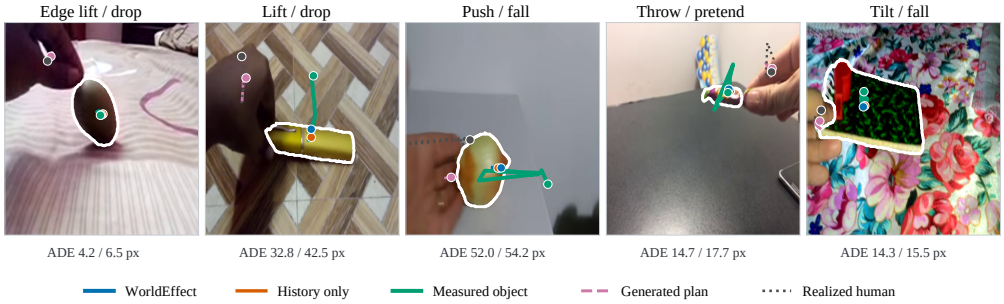


Figure 11: **World response from generated human plans.** Real last-observed RGB is a rendering background only. Blue, orange, and green trace WorldEffect, history-only, and measured object motion; dashed magenta and dotted gray trace generated and realized wrist motion. Within each action family, we retain cases at or above median positive response gain and display the one with lowest plan ADE. Aggregate evaluation above includes all 69 accepted videos across both roles.

History-only retraining measures the value of conditioning after learning; frozen plan removal tests input dependence. MMHOI's 585-window final evaluation covers twenty scenarios, separately from its 563-window validation diagnostic: Transformer/WorldEffect ADE is 24.98/22.14 cm. HOI-M3 gives 7.821/4.504 cm on three final sequences, which define its independent units.

B.6 Interaction Charts and Egocentric Transfer

Temporal alignment in camera space. Figure 12 uses measured motion, visibility, and geometry to select cases, without reading prediction errors. Across all 2,590 eligible Aria windows, alignment reduces normalized projection error (NPE) from 0.3289 to 0.2590 and pixel error from 119.49 to 90.60, while IoU increases from 0.4554 to 0.5167. Unseen-category values improve from 0.2735/77.16/0.4921 to 0.2358/66.05/0.5239. NPE uses the measured box diagonal.

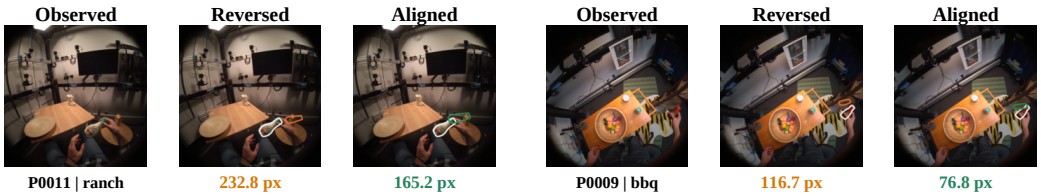


Figure 12: **Aligned plans improve measured Aria projections.** White is the measured amodal outline; orange and green project reversed- and aligned-plan futures. Case selection uses measured motion, visibility, and geometry; cohort comparisons use all 2,590 eligible windows.

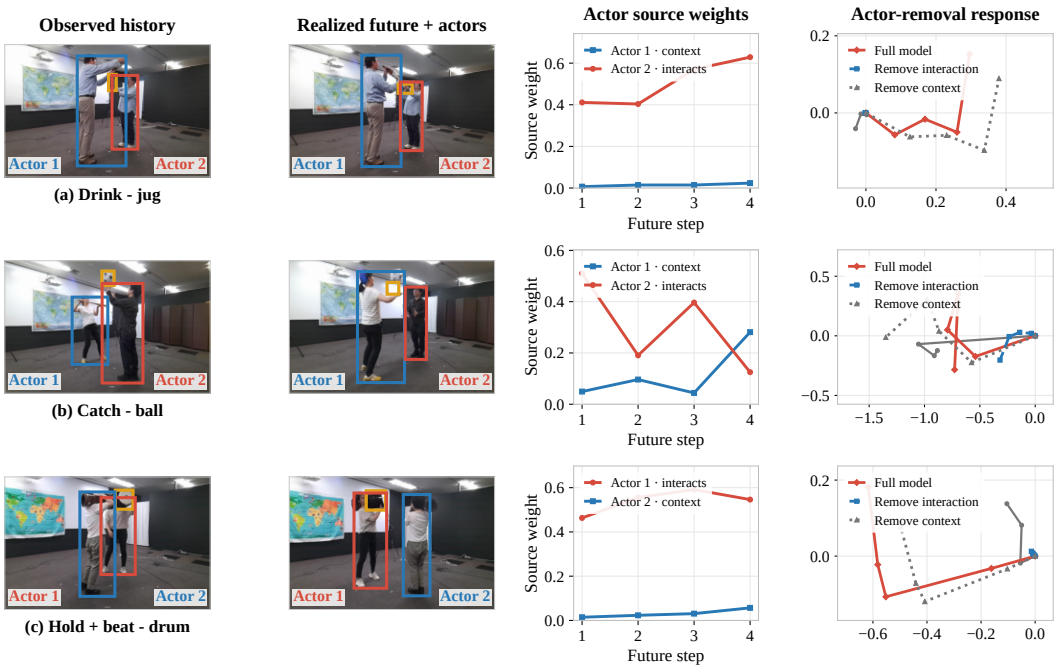
With the same frozen experts, reliability/compatibility/factorized inference gives identity-equal Joint ADE 7.68/7.54/7.48 on Quest 3 and 6.56/6.44/6.41 on Aria in frozen transfer. Archive-disjoint results are 6.42/6.29/6.24 and 7.09/6.97/6.93, respectively. Factorization adds a smaller gain to compatibility in both transfer roles, with unchanged expert predictions.

A 3,974-parameter router gives joint errors 6.27/6.85 versus analytic 6.24/6.93 and fixed 6.62/7.21. Learned–analytic difference intervals include zero; experts and forward counts match. Clip-disjoint evaluation shares 54 development capture IDs; only learned routing requires fitting.

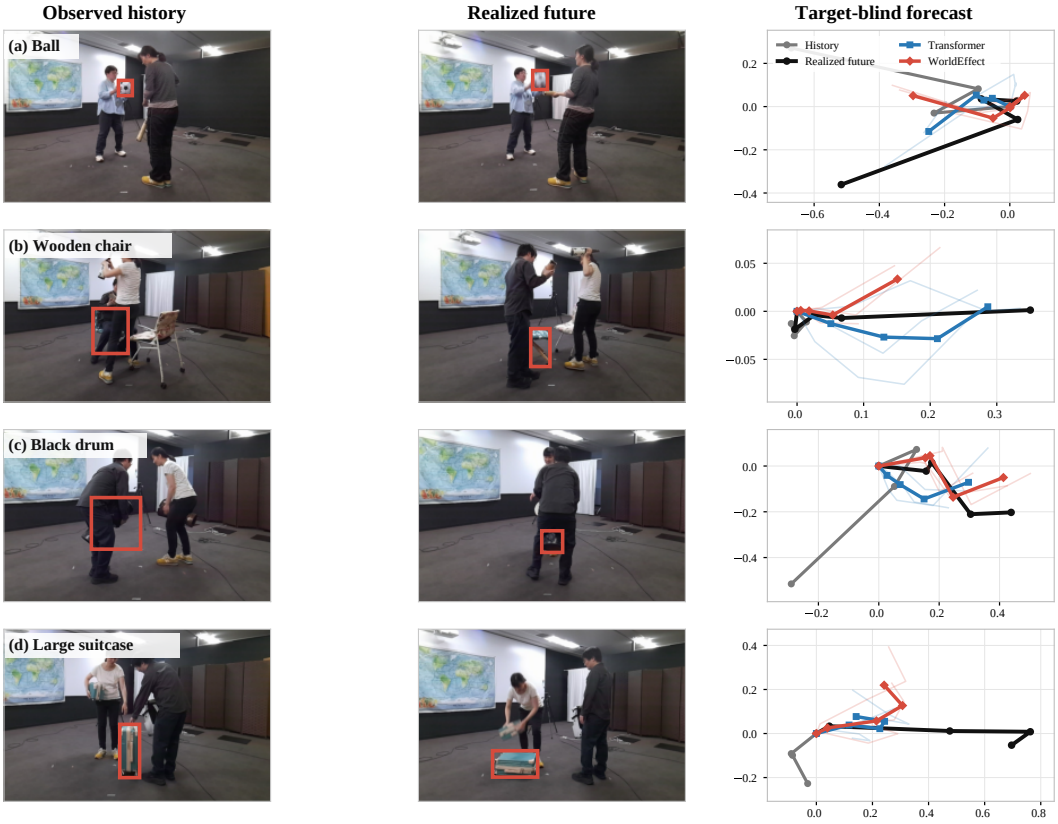
The reversal diagnostic reverses absolute future wrist samples. It preserves the sample set but changes endpoints and boundary continuity, measuring temporal-alignment sensitivity. In the leave-one-person-out study, response heads are retrained for each of 13 identities while chart specialists remain frozen. Identity-equal joint reductions are 3.73% ($[2.92, 4.53]\%$) and 3.72% ($[2.27, 4.81]\%$) for the two reported evaluation roles, with equal weight per identity.

B.7 Qualitative Response and Actor Attribution

Figure 13 contrasts removal of the annotated actor versus a context actor, conditional on the supplied action edge. Lower plots use target-motion-selected cases: gray denotes history, blue the plan Transformer, red WorldEffect predictions, and black the recorded object trajectory.



(a) Conditional actor attribution



(b) Recorded interactions and predicted trajectories

Figure 13: **MMHOI attribution and trajectories.** (a) Red/blue/yellow identify the interacting actor, context actor, and object; source weights and removal effects follow the annotated actor. (b) Ball, chair, drum, and suitcase cases use target-motion proximity to task medians. Thin/bold curves denote seeds/means; black is measured motion, blue the Transformer, and red WorldEffect.

C Discussion and Future Directions

Action evidence across sources. The experiments connect three forms of action evidence. Shared-history constructed pairs isolate a change in the supplied plan, while recorded interactions test response prediction under measured motion. Same-state simulation supplies independently executed outcomes for both alternatives. Together, these settings show that supplied human motion helps determine object response and that paired supervision improves response error in the controlled construction. Their differences are also informative: ordinary state supervision nearly matches the paired objective in the six-second simulation, whereas the constructed-pair study reveals a larger learning effect. A useful next step is measured human execution with matched initial states, spanning near misses, light contact, and sustained manipulation, to connect these two regimes.

Coordinates, coverage, and decisions. Coordinate compatibility allocates trust using plan geometry, and the candidate cover expresses remaining response variation. Same-expert chart comparisons isolate the weighting rule; same-base residual predictors isolate candidate construction. Refitted component controls additionally measure how downstream adaptation changes the contribution of a fixed inference rule. These comparisons connect coordinate reasoning to uncertainty representation without conflating expert quality, candidate coverage, and probability assignment. The joint-risk and energy-score evaluations measure complementary properties of the issued distribution, while executable decisions and calibration receive separate tests. Future work can optimize task-specific decisions and calibrate confidence on independent sources.

Interaction-relevant input quality. Generated-plan results identify hand position, orientation, and timing as important aspects of plan quality. Evaluating upstream motion and downstream response together identifies the features that matter for interaction: wrist trajectories and orientation complement global root accuracy. Training with out-of-fold generated plans can expose the response model to realistic upstream errors while preserving an independent evaluation. For video-derived states, quality filtering changes the action distribution of evaluable clips. Reporting extraction coverage alongside conditional prediction error makes this input-quality dependence visible.

Geometric validity and longer interaction. The present output is a structured rigid-object response. Valid rotations preserve pose geometry, while contact, support, and force consistency require additional evidence. Boundary edges in 15 of 18 object templates make signed penetration measurements sensitive to mesh quality; validated surface geometry and contact annotations are priorities for stronger physical diagnostics. Longer interaction calls for state memory and response feedback that update supplied plans during execution. A closed-loop evaluation can alternate action execution, state observation, and replanning, measuring both response accuracy and task completion. Extending the state to articulated and deformable objects would test the same compatibility principle with richer interaction constraints and geometry.