CALM-VR: CANDIDATE-SPECIFIC TEMPORAL READOUT FOR EARLY VR TARGET PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Early target prediction in virtual reality (VR) ranks changing legal objects from causal multimodal histories. We ask whether candidate-specific temporal readout adds value when candidate information already enters both the final scorer and aligned history summaries. **CALM-VR** encodes gaze, head, and controller histories with modality-specific GRUs, then uses each candidate token to retrieve temporal evidence for masked ranking. A parameter-matched broadcast-query control shares one readout while retaining individual candidate tokens and aligned summaries; crossed controls vary summary alignment separately. An event-aligned, three-task collection provides mutually exclusive 97/21/30 fitting/selection, development/calibration, and frozen-protocol evaluation roles. In the 30-participant evaluation, 500-ms Top-1 was **51.20%** versus **46.42%** for designated late fusion ($+4.78$ pp, 95% CI $[2.88, 6.68]$, Holm $p < .001$). The prespecified secondary query comparison estimated a $+2.40$ pp overall gain (95% CI $[1.35, 3.45]$, Holm–3 $p = .027$). Equal task weighting gave a descriptive $+1.40$ pp point estimate; task-specific contrasts ranged from -3.56 pp in target acquisition to $+6.15$ pp in sequential assembly. Smaller positive aggregate gains were observed at 800 and 1000 ms. These results quantify the incremental value of candidate-specific temporal readout beyond candidate-conditioned scoring and aligned summaries, with observed variation across tasks in the evaluated VR setting.

1 INTRODUCTION

Eye, head, controller, and scene-state signals begin to organize before a user presses a trigger or grabs an object. Predicting the next target from this causal history could allow a VR system to prefetch an asset, initialize a grasp constraint, or prepare task-specific content while leaving the final choice to the user. Prior studies show that gaze dynamics indicate impending interaction (David-John et al., 2021; Narkar et al., 2024), and that eye–head endpoint distributions improve target selection in head-worn displays (Wei et al., 2023). The deployment setting, however, is harder than recognizing an action after it occurs.

Target-conditioned history modeling is established in recommendation: DIN constructs ad-dependent representations of behavioral history (Zhou et al., 2018), while DIEN models target-relative interest evolution (Zhou et al., 2019). Building on this line, we study the incremental value of candidate-specific temporal readout for early VR target ranking when candidate information already enters both the final scorer and candidate-aligned history summaries. Our contribution is a controlled empirical characterization of this design choice over continuous multimodal histories and changing legal object sets, rather than a new attention operator. Parameter-matched and crossed-summary comparisons separate the readout choice from these other candidate-information pathways.

CALM-VR implements candidate-specific temporal readout after modality-specific GRU encoding (Figure 1). The broadcast-query control replaces only the temporal query with the mean legal-candidate token, retaining each individual token in the final scorer. Both architectures therefore use candidate information; the comparison tests the value of reading temporal evidence separately for each candidate. Consumer eye tracking can drift across users or after re-donning (Aziz et al., 2024; Bozkir et al., 2023), while uncertain actions require calibrated probabilities (Guo et al., 2017). Reliability-aware fusion, bounded gaze-bias updates from completed interactions (Hou et al., 2025),

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.

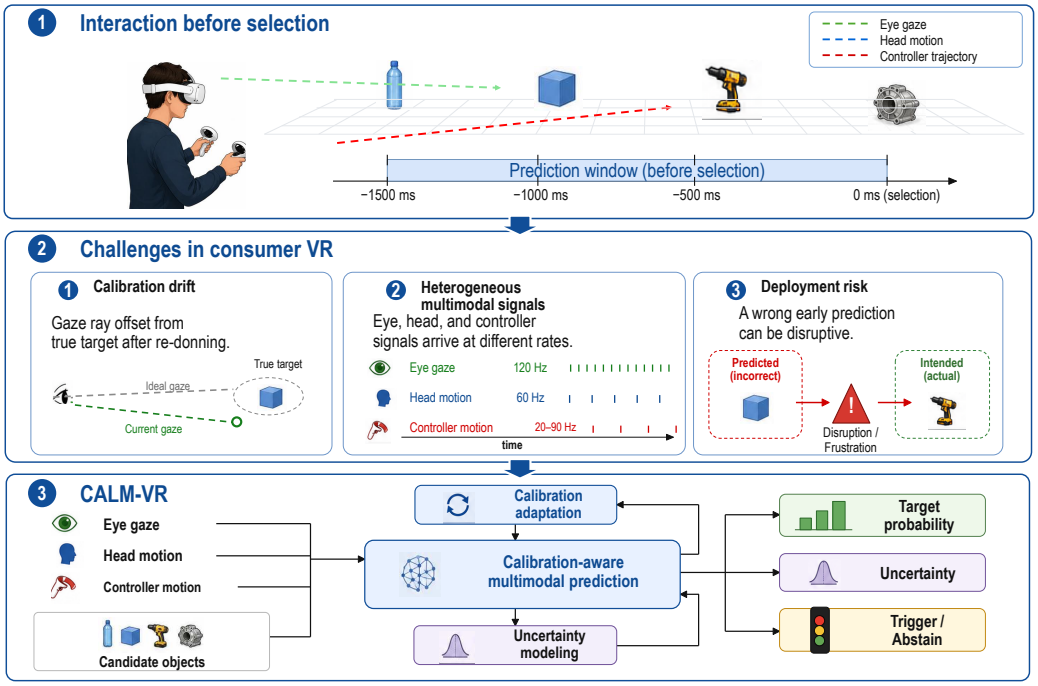

Figure 1: CALM-VR overview. Each candidate reads temporal evidence from GRU-encoded multimodal histories before shared masked scoring. Bounded gaze correction, reliability-aware fusion, and a frozen adaptive-prediction-set policy provide deployment support.

temperature scaling, and conformal sets (Romano et al., 2020; Angelopoulos et al., 2021) serve as deployment components, not additional algorithmic contributions.

Our contributions are: (1) a controlled study of the incremental value of candidate-specific temporal readout beyond candidate-conditioned scoring and aligned history summaries in early VR target ranking; (2) an exact parameter-matched comparison and crossed-summary, task, and horizon analyses that characterize both the benefits and the boundaries of this design choice; and (3) an event-aligned multimodal collection with explicit participant roles and a frozen evaluation protocol, supporting transfer and offline deployment analyses within the documented platform.

2 Problem Formulation

Each trial ends at selection time T with target $Y \in \mathcal{O}_T$. At horizon Δ, the anchor is $t = T - \Delta$ and the available information is

$$X_t = \{G_{t-W:t}, H_{t-W:t}, C^L_{t-W:t}, C^R_{t-W:t}, Q_{t-W:t}\},$$
$$\mathcal{F}_t = \sigma\{X_u, \mathcal{O}_u : u \le t\} \vee \sigma\{Y_i : T_i < t\}, \tag{1}$$

where G, H, C^L, C^R are gaze, HMD, and controller streams, and Q stores time and validity. The legal set $\mathcal{O}_t$ is fixed by interaction state at the anchor; an occluded but selectable object may remain legal, while a visible object from a later assembly step is excluded. We retain only events with $\mathcal{O}_t \neq \varnothing$ and $Y \in \mathcal{O}_t$; hence the denominator below is positive and the supervised target lies in the model support. With padded slots $j \le N_{\max}$, mask $m_j = \mathbf{1}[o_j \in \mathcal{O}_t]$, and score s_j, the model distribution is

$$p_\theta(j \mid \mathcal{F}_t, \Delta) = \frac{m_j e^{s_j}}{\sum_{k=1}^{N_{\max}} m_k e^{s_k}}. \tag{2}$$

Thus invalid slots have zero mass, legal probabilities sum to one, and adding masked padding cannot change them. All histories and candidate summaries are constructed at or before t, so the forward map is $\mathcal{F}_t$-measurable: two trials identical through the anchor receive identical predictions regardless

of later records. This is a computational non-anticipation guarantee, not an independence assumption. We evaluate participant-equal Top-1 accuracy at $\Delta \in \{500, 800, 1000\}$ ms; 500 ms is the principal endpoint.

Let q_j be the token of legal candidate j, $\bar{q} = |\mathcal{O}_t|^{-1} \sum_{j \in \mathcal{O}_t} q_j$ its set mean, Z the modality-specific encoded histories, and e_Δ the horizon token. Writing $\mathcal{R}$ for temporal retrieval and fusion and f for the shared scorer, we compare

$$\begin{aligned} \text{Shared readout:} \quad & s_j = f(q_j, \mathcal{R}(\bar{q}, Z), e_\Delta)\,, \\ \text{Candidate-specific readout:} \quad & s_j = f(q_j, \mathcal{R}(q_j, Z), e_\Delta)\,. \end{aligned} \tag{3}$$

Both architectures condition the final score on the candidate, and both can use candidate-aligned input summaries. The distinction is whether temporal evidence is read out once and shared, or retrieved separately for each candidate. We study whether the latter provides a useful inductive bias beyond candidate-conditioned final scoring and aligned summaries. A shared readout supplies the same retrieved evidence to every score, even though the scorer can still distinguish candidates through q_j. Candidate-specific readout additionally lets each alternative select different time samples from the GRU-encoded histories. This separates where evidence is retrieved from how candidate-conditioned scores are computed.

3 CALM-VR

Figure 2 summarizes the pipeline. Candidate-specific temporal readout implements the design choice in Eq. (3); adaptation and selective action wrap the frozen predictor for deployment.

3.1 Candidate-Specific Temporal Readout

For candidate center $p_{j,t}$ and HMD pose $(p_{H,t}, R_{H,t})$, we compute

$$r_{j,t} = R_{H,t}^\top (p_{j,t} - p_{H,t}). \tag{4}$$

The 49-dimensional candidate vector c_j excludes raw world position and contains HMD-relative geometry, state, and causal alignments of gaze, head, controller rays, and controller velocity. Let $\mathbf{C}_{\text{obj}} = (c_1, \ldots, c_{N_{\max}})^\top$ denote the padded candidate matrix. A shared MLP gives $q_j = \phi_\theta(c_j) \in \mathbb{R}^{192}$. Separate two-layer GRUs (Cho et al., 2014) encode 91 samples over 1500 ms as $Z^m = (z_1^m, \ldots, z_S^m)$ for $m \in \{G, H, L, R\}$. Invalid sensor payloads are zeroed while their validity/time metadata remain; only history padding is attention-masked. For attention head a ($d_a = 48$) (Vaswani et al., 2017),

$$\begin{aligned} \omega_{js}^{m,a} &= \operatorname*{softmax}_{s:\eta_s=1} \left(\frac{(W_Q^{m,a} q_j)^\top W_K^{m,a} z_s^m}{\sqrt{d_a}} \right), \\ h_j^m &= W_O^m \operatorname*{\overset{4}{concat}}_{a=1} \sum_{s:\eta_s=1} \omega_{js}^{m,a} W_V^{m,a} z_s^m. \end{aligned} \tag{5}$$

Let u^m collect the modality-valid fraction over all history samples, validity at the latest sample, and valid fraction over the latest six cached samples. With horizon token e_Δ, the event-level router and shared scorer are

$$\begin{aligned} \rho &= \operatorname{softmax} g_\theta([u^G; u^H; u^L; u^R; e_\Delta]), \\ h_j &= \sum_m \rho^m h_j^m, \qquad s_j = f_\theta([q_j; h_j; e_\Delta]). \end{aligned} \tag{6}$$

Because $\rho^m \geq 0$ and $\sum_m \rho^m = 1$, h_j lies in the convex hull of the modality-specific evidence and, by the triangle inequality, $\|h_j\| \leq \max_m \|h_j^m\|$ for any norm. The router thus combines the four retrieved representations with bounded, validity-dependent weights.

Proposition 1 (candidate-coordinate invariance and set equivariance). The HMD-relative candidate feature map is invariant to a joint global rigid change of its positions and directions. Moreover, for any permutation matrix P applied jointly to candidate rows and their mask, the evaluation map satisfies $s(P\mathbf{C}_{\text{obj}}) = Ps(\mathbf{C}_{\text{obj}})$ and $p(P\mathbf{C}_{\text{obj}}, Pm) = Pp(\mathbf{C}_{\text{obj}}, m)$.

Relative coordinates preserve the candidate features under joint rigid transforms, and shared row-wise operations preserve object scores under reordering; the proof is in Appendix E.1.

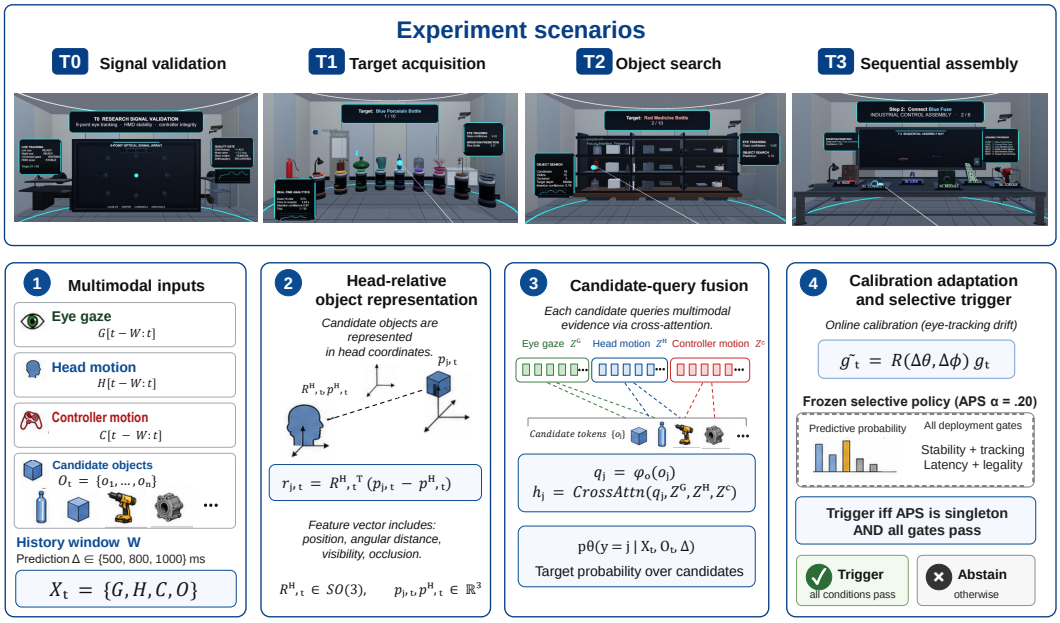

Figure 2: CALM-VR pipeline and experimental setting. The top row summarizes T0 signal validation and the T1–T3 prediction tasks. Modality-specific GRUs encode causal histories; HMD-relative candidate tokens retrieve temporal evidence for shared scoring at 500, 800, and 1000 ms. Completed selections update a two-angle gaze bias. The frozen selective policy triggers only when the APS set at $\alpha = .20$ is a singleton and all stability, tracking, latency, and legality gates pass; otherwise it abstains.

3.2 Online Two-Angle Adaptation

After completed selection k, we compute the clipped yaw–pitch discrepancy between observed gaze g_k and the selected-object direction d_{y_k} and update the bounded bias:

$$\begin{aligned} e_k &= \mathrm{clip}\big(a(d_{y_k}) - a(g_k), [-b, b]^2\big)\,, \\ \delta_k &= \gamma\delta_{k-1} + (1-\gamma)e_k, \end{aligned} \tag{7}$$

where $a(\cdot)$ maps an HMD-frame direction to yaw–pitch angles, $\gamma = .6$, $b = 3.5°$, and $\delta_0 = 0$. The updated state is available only from trial $k+1$, corrects valid gaze-history vectors and their candidate-alignment fields, and leaves network weights frozen. This recursion guarantees $\|\delta_k\|_\infty \le b$ and is causal because only completed selections contribute to the state (Proposition 2, Appendix E.3). The counterfactual Unity-policy replay uses a separate bounded projected optimizer described in Appendix E.3.

3.3 Training and Selective Output

For legal-set probability vector p and target y, training uses

$$\ell_\lambda(p, y) = -\log p_y + \lambda \sum_{j \in \mathcal{O}} (p_j - \mathbf{1}[j = y])^2, \qquad \lambda = .1. \tag{8}$$

The 500-ms primary loss receives weight 2.0 and each 800/1000-ms auxiliary loss receives weight 0.5. The frozen 2.584M-parameter network has width 192, hidden-layer dropout .1, modality dropout 0, and no semantic class input; grouped cross-validation (CV) fixed six epochs before any access to the 30-person data. AdamW (Loshchilov & Hutter, 2019), the Brier term (Brier, 1950), and remaining constants are reported in the supplement.

The masked cross-entropy plus Brier objective is a strictly proper scoring rule; the complete statement and proof are in Appendix E.2.

Temperature scaling and APS (Romano et al., 2020) are fit on disjoint groups of 10 and 11 development participants, respectively, without evaluation-cohort access. Because pooled event-level

Table 1: Data accounting and frozen protocol.

Item	Value
Original completed cohort	150 participants
QC-passed sessions	139 participants
Model-selection CV	97; participant-grouped 5-fold
Post-hoc development	21; five fixed seeds
Legacy QC-only partition	21; descriptive summaries only
QC exclusions	11; protocol-defined analytic QC
Model-unseen cohort	30 participants
Principal endpoint	500-ms participant-equal Top-1
Secondary horizons	800 and 1000 ms

exchangeability is not established for repeated, task-ordered observations (Dunn et al., 2023), APS is a frozen empirical set-valued interface, with no claimed finite-sample conformal validity. At $\alpha = .20$, the policy acts only on a singleton prediction set that passes the frozen stability, tracking, latency, and legality gates in Figure 2; otherwise it abstains. The construction, operating constants, and conditional guarantee under exchangeability appear in Appendix E.3.

4 Experimental Setup

4.1 Participants, Tasks, and Data Splits

The original collection comprised 150 participants, of whom 139 passed analytic QC. Mutually exclusive roles assigned 97 participants to fitting/grouped CV, 21 to post-hoc development and calibration, and 21 to a previously opened legacy partition used only for descriptive QC. A separately acquired 30-participant cohort was kept unopened until the predictive protocol was frozen and then used only for offline evaluation and counterfactual policy analysis (Table 1). The original and evaluation cohorts comprised 75/75 and 15/15 women/men, respectively; across cohorts, age was $M = 25.1$ years (SD $= 5.6$). All participants provided informed consent, met vision and VR-safety criteria, received fixed compensation, and were covered by institutional ethics approval.

Sessions used a Meta Quest Pro, two Touch Pro controllers, and a PC-based Unity application. T0 calibration required at least 80% combined-gaze validity and mean fused angular error below 2°. T1, T2, and T3 comprised target acquisition, cluttered-shelf search, and six-step assembly, respectively. The protocol specifies 38 target-bearing selections (10+10+18). The 500-ms development caches retain 38 target labels per modeled participant; 23 of the 139 analytic raw logs contain 39–40 valid selection records because retry or recovery records are preserved for QC. Each prediction used 91 samples from a 1500-ms pre-anchor history and at most 25 legal candidates. A monotonic clock linked sensor samples, candidate snapshots, responses, and executed actions. Events were discarded when the realized horizon differed from nominal by more than 20 ms, the selected target was absent from the anchor-time legal set, or that set was invalid. The frozen 30-person evaluation retained 1,131/1,123/1,114 events at 500/800/1000 ms; decreasing counts reflect horizon-specific causal-history coverage and protocol exclusions, with no imputation.

4.2 Baselines, Metrics, and Statistics

We compared CALM-VR with a geometric gaze-ray method, a head-plus-controller (H+C) model, and gaze–head–controller (GHC) rankers based on LightGBM, GRU, TCN, a late-fusion Transformer, and a mixture of experts (MoE). All GHC candidate rankers used the same causal cache, modalities, candidate features, and anchor-time legal mask. Baseline configurations appear in Table S3. The neural GHC baselines condition final candidate scoring on candidate features but do not perform per-candidate cross-attention into each modality-specific history before fusion. The late-fusion comparison therefore measures the complete configuration difference, whereas the broadcast-query control isolates temporal querying. The late-fusion Transformer–GHC was the strongest designated model at the principal 500-ms endpoint. Participant and session identifiers, scene labels, absolute world centers, post-anchor samples, and realized targets were excluded from model inputs.

The principal endpoint was mean per-participant Top-1 accuracy at 500 ms; 800 and 1000 ms were secondary horizons. For participant p with n_{ph} valid events at horizon h, we computed

$$a_{ph} = \frac{1}{n_{ph}} \sum_{e=1}^{n_{ph}} \mathbf{1}[\widehat{y}_{phe} = y_{phe}], \qquad A_h = \frac{1}{|\mathcal{P}_h|} \sum_{p \in \mathcal{P}_h} a_{ph}. \tag{9}$$

where $\mathcal{P}_h$ contains participants with $n_{ph} > 0$. Participants receive equal inferential weight. Predictive contrasts use 50,000 participant-paired bootstrap resamples for 95% percentile CIs and 50,000 sign-flip draws for two-sided tests (Appendix L). Holm correction applies separately to the five complete-system comparisons and the three query horizons. All five complete-system tests have adjusted $p < .001$. Counterfactual-policy families remain separate. Grouped CV and five-seed post-hoc analyses provide development evidence. Task-macro Top-1 first computes accuracy within participant and task, then averages T1–T3 equally. Random-choice accuracy uses each event's realized legal-set size, and ranking quality is summarized by MRR and chance-normalized MRR. Frozen prediction, secondary transfer, and counterfactual policy results retain their respective analysis roles.

Model identity, data access, and freezing. Grouped CV on the 97-person fitting partition selected the architecture and six-epoch rule. Before opening any evaluation-cohort record, we froze model and baseline definitions, features, hyperparameters, training budgets, checkpoint selection, endpoints, comparators, metrics, statistical tests, and the selective policy. The evaluation cohort never entered fitting, calibration, model selection, or protocol selection. The primary predictive comparison was CALM-VR versus designated late fusion at 500 ms. Full versus broadcast query was a prespecified secondary architectural analysis: both variants, their training and checkpoint-selection rules, the three horizon contrasts, and the separate Holm–3 family were fixed before evaluation-cohort access. Secondary status denotes the analysis hierarchy, not post hoc comparator selection.

5 Experimental Results

5.1 Frozen 30-Participant Offline Evaluation

Table 2 and Figure 3 report the frozen offline evaluation in previously unseen participants. At 500 ms, CALM-VR achieved 51.20% Top-1 versus 46.42% for the designated late-fusion Transformer–GHC comparator ($+4.78$ pp, 95% CI $[2.88, 6.68]$, Holm $p < .001$). Gains over H+C and gaze ray were 19.6 pp ($[16.5, 22.7]$) and 11.6 pp ($[9.1, 14.1]$). At 800 and 1000 ms, CALM-VR reached 19.0% and 19.1%, exceeding the same late-fusion comparator by $+5.6$ pp (CI $[4.2, 7.0]$) and $+5.9$ pp ($[4.4, 7.4]$), respectively. All five predictive contrasts have Holm-adjusted $p < .001$. The principal endpoint remains 500 ms, but the improvement is consistent across all three horizons.

Task-level estimates. The 500-ms full-minus-broadcast point estimates were -3.56 pp in T1 target acquisition, $+1.60$ pp in T2 cluttered-shelf search, and $+6.15$ pp in T3 sequential assembly (Table S7). Equal task weighting reduced the aggregate query point estimate from $+2.40$ to a descriptive $+1.40$ pp (51.44% versus 50.04%), reflecting the larger number of assembly events in the 10:10:18 protocol. Both metrics first aggregate within participant, using retained events, and then weight participants equally. The observed query contrasts varied in direction and magnitude across tasks, with the largest point estimate in six-candidate assembly rather than 25-candidate search. These task summaries are descriptive; participant-level inference is reported for the prespecified aggregate horizon contrasts in Table 3. Against late fusion, equal-task accuracy was 51.44% versus 46.59% ($+4.85$ pp, 95% CI $[2.70, 7.00]$).

Candidate-set sensitivity. On the same 297 T2 events, a nested candidate-set sweep increased the full-minus-broadcast gap from 0.3 pp at $K = 2$ to 1.6 pp at $K = 25$. Subsets retain the target and geometrically nearest distractors from the original legal set. The $K = 25$ endpoint recovers the original T2 condition: 56.0/54.4% for full/broadcast query, 50.0% for late fusion, and 4.0% chance (Table S8).

Zero-shot transfer. Separate transfer trials from the same 30 evaluation participants tested held-out layouts and two unseen task flows without retraining or adaptation (Table S9). The layout condition

Table 2: Frozen 30-person evaluation (Top-1, %). Gains are pp [participant 95% CI]; query tests use a separate Holm–3 family. Temperature scaling preserves Top-1. Broadcast-minus-late-fusion gaps are descriptive.

Model / contrast	**500 ms**	**800 ms**	**1000 ms**
Gaze ray	39.60	4.00	3.80
H+C	31.60	12.50	12.00
LightGBM–GHC	42.50	12.60	13.00
GRU–GHC	42.80	12.50	12.80
TCN–GHC	43.00	12.70	13.10
Late-fusion Trans.–GHC	46.42	13.40	13.20
MoE–GHC	43.20	13.20	13.30
CALM-VR, before temp. scaling	51.20	19.00	19.10
Broadcast query	48.80	18.10	18.20
CALM-VR, after temp. scaling	**51.20**	**19.00**	**19.10**
Δ vs. late-fusion GHC	+4.78 [2.88,6.68]	+5.60 [4.20,7.00]	+5.90 [4.40,7.40]
Holm p	$< .001$	$< .001$	$< .001$
Query gain: full − broadcast	+2.40 [1.35,3.45]	+0.90 [0.15,1.65]	+0.90 [0.13,1.67]
Holm–3 p	.027	.036	.036
Broadcast − late fusion (pp)	+2.38	+4.70	+5.00

Table 3: Parameter-matched candidate-query evidence: full-minus-broadcast Top-1 gains in pp [participant 95% CI]. Development p values are unadjusted sign-flip tests; evaluation p values use the separate Holm–3 family.

Partition	**500 ms**	**800 ms**	**1000 ms**
97-person OOF	**+2.41** [1.58,3.24]	**+0.92** [0.39,1.45]	**+0.88** [0.35,1.41]
Sign-flip p	.0012	$\approx .002$	$\approx .003$
21-person validation	**+2.36** [1.24,3.46]	**+0.86** [0.10,1.62]	**+0.84** [0.10,1.58]
Sign-flip p	.021	$\approx .03$	$\approx .03$
30-person evaluation	**+2.40** [1.35,3.45]	**+0.90** [0.15,1.65]	**+0.90** [0.13,1.67]
Holm–3 p	.027	.036	.036

retained seen task logic and 1,128 events: full/broadcast Top-1 was 48.20/46.60%, giving a $+1.60$ pp query increment (95% CI $[0.45, 2.75]$). U1/U2 retained 596/594 events with legal-set sizes 12/18. Their task-macro accuracy was 44.65% for full query, 42.75% for broadcast query, and 39.65% for late fusion. The task-macro query increment was $+1.90$ pp ($[0.75, 3.05]$). Complete-system differences over late fusion were $+4.90$ pp on held-out layouts and $+5.00$ pp across unseen tasks. These secondary estimates support candidate-specific retrieval across layout and task-flow changes within the evaluated platform.

5.2 Parameter-Matched Readout Analysis

We compared full and broadcast-query models with identical parameter counts to estimate the incremental value of candidate-to-history assignment. Full CALM-VR uses the embedded candidate token $q_j = \phi_\theta(c_j)$ in each temporal cross-attention block and candidate-aligned history summaries. The matched control replaces only the query with the masked candidate-set mean $\bar{q}$, broadcast to every candidate; the individual q_j still enters the shared scorer. Both variants retain the same cache, folds, temporal encoders, attention, router, scorer, loss, training budget, and exactly 2,583,749 parameters. The supplement retains the crossed history-summary control and reports the ordered bridge analysis.

The OOF readout analysis uses grouped-CV predictions from the reference seed, while validation averages five fixed seeds. The 21-person partition remains post-hoc development evidence, and the 30-person comparison is a prespecified secondary architectural analysis. Table 3 reports the corresponding participant-level intervals and distinguishes unadjusted development tests from the prespecified evaluation Holm–3 family.

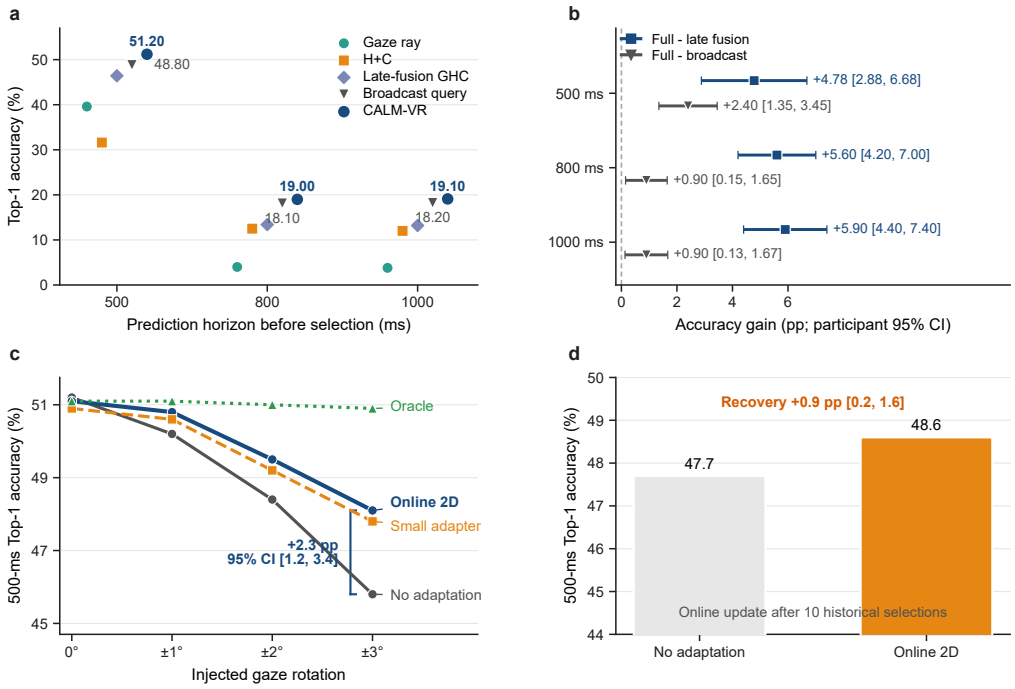

Figure 3: Frozen offline performance and robustness in 30 previously unseen participants. (a) Participant-equal Top-1 accuracy across horizons, including the parameter-matched broadcast-query control. (b) Complete-system gains over designated late-fusion Transformer–GHC and query gains over broadcast query; all participant-level 95% intervals exclude zero. The complete-system tests have Holm $p < .001$; query tests use the separate Holm–3 horizon family ($p = .027, .036, .036$). (c) Accuracy under injected gaze rotation before and after ten historical selections. (d) Absolute 500-ms Top-1 accuracy before and after online adaptation following physical re-donning. Where shown, error bars are 95% intervals; oracle correction is an upper bound.

The full-minus-broadcast effect was largest at the principal 500-ms endpoint and replicated in 97-person OOF ($+2.41$ pp, CI $[1.58, 3.24]$), 21-person validation ($+2.36$ pp, $[1.24, 3.46]$), and the frozen 30-person evaluation ($+2.40$ pp, $[1.35, 3.45]$, Holm–3 $p = .027$). The corresponding gains at 800 ms were $+0.92$, $+0.86$, and $+0.90$ pp, and those at 1000 ms were $+0.88$, $+0.84$, and $+0.90$ pp; all nine participant-level intervals exclude zero (Table 3). In the 30-person evaluation, the 800/1000-ms intervals were $[0.15, 1.65]$ and $[0.13, 1.67]$, with both Holm–3 $p = .036$. Thus, candidate-specific temporal readout improves the aggregate result at each evaluated horizon, with a substantially larger increment at 500 ms.

The descriptive ordered bridge and query/full-gap ratios are reported in Appendix J.2 and Table S15; their increments depend on the displayed comparison path.

Crossed history-summary control. The parameter-matched 2×2 analysis crosses temporal-query identity with alignment of the 36-D history summaries (Table S12). At 500 ms, query gains with aligned versus set-mean summaries are 2.41 versus 1.47 pp in OOF and 2.36 versus 1.40 pp in validation. The difference of query gains is 0.95 pp in each partition. Candidate-specific readout thus adds value under either summary representation, with a larger increment when summaries remain aligned. At 800/1000 ms, query gains remain $0.92/0.88$ pp in OOF and $0.86/0.84$ pp in validation under either representation. This descriptive interaction analysis holds all four configurations at exactly 2,583,749 parameters.

Parameter-matched set-aware baselines. At 500 ms, Set Transformer–GHC and Deep Sets–GHC use the same causal cache, candidate features, legal mask, splits, and training budget, with parameter counts within $\pm 5\%$ of CALM-VR (Table S16). CALM-VR exceeds Set Transformer by $+4.27$ pp

Table 4: Participant-equal counterfactual policy means ($n = 30$). Correct/waste rates are conditional on triggers; starred endpoints are policy-defined.

Policy	Wait (ms)	Correct (%)	Waste (%)	Wrong/min	Trigger (%)
No assist	1,418.0	–	–	0^*	0^*
Always top-1	1,219.6	51.1	48.9	1.28	100^*
Selective CALM-VR	1,388.5	96.9	3.1	0.01	6.5

in OOF and $+4.15\,\mathrm{pp}$ in validation (95% CIs $[3.38, 5.18]$ and $[3.03, 5.25]$); the corresponding gaps over Deep Sets are $+7.11/+7.06\,\mathrm{pp}$ ($[6.05, 8.19]/[5.62, 8.48]$). These development comparisons support the value of temporal candidate-to-history binding beyond parameter count and unordered-set handling. They complement, rather than replace, the exact parameter-matched broadcast-query control.

5.3 Counterfactual Utility–Risk Trade-off

We applied frozen No-assist, Always top-1, and Selective CALM-VR policies offline to the 30-person records. Policy-specific validity filtering retained 1,135, 1,132, and 1,131 opportunities. Estimated waiting spans selection confirmation to resource readiness; correct/waste rates are conditional on counterfactual triggers, trigger rate is per valid opportunity, and inference uses participant-level aggregates.

Selective CALM-VR triggered on 6.5% of opportunities, with 96.9% correct and 3.1% wasted prefetches among triggers. Relative to No-assist, it reduced estimated waiting by 29.5 ms (95% CI $[-50.6, -8.4]$), while yielding 0.01 incorrect triggers/min. Full paired contrasts appear in Table S11. Development-set calibration diagnostics (Appendix F.3) distinguish confidence ranking, which yielded lower selective risk at matched action rates, from APS, which defines the frozen set-valued action interface.

5.4 Robustness and Runtime

CALM-VR was most sensitive to gaze corruption: missing gaze and a 300-ms gaze gap reduced 500-ms Top-1 by 18.4 and 11.4 pp, respectively, whereas controller and temporal-shift perturbations had smaller effects (Table S17). Under $\pm 3^\circ$ synthetic gaze drift, online correction improved accuracy from 45.8% to 48.1% after ten completed selections ($+2.3\,\mathrm{pp}$, 95% CI $[1.2, 3.4]$); physical re-donning improved accuracy from 47.7% to 48.6% ($+0.9\,\mathrm{pp}$, $[0.2, 1.6]$). Full drift trajectories and oracle comparisons appear in Tables S18–S19.

CALM-VR has 2.584M parameters and batch-one RTX 4080 SUPER GPU P50/P95 inference latency of 12.54/12.62 ms, excluding Unity and end-to-end system overhead (Table S20). Additional calibration and selective-risk diagnostics are reported in Appendix F.3.

6 Conclusion

CALM-VR studies whether each candidate benefits from its own temporal readout after modality-specific history encoding. The frozen complete system reached 51.20% Top-1 versus 46.42% for late fusion at 500 ms. The prespecified secondary parameter-matched analysis estimates a $2.40\,\mathrm{pp}$ aggregate readout gain, with smaller positive aggregate gains at 800/1000 ms. Descriptive equal task weighting gives a $1.40\,\mathrm{pp}$ point estimate. Task-specific point estimates were negative in T1 and positive in T2/T3. Together with the crossed-summary controls, these findings quantify the incremental value of candidate-specific readout beyond candidate-conditioned scoring and aligned summaries, with observed variation across tasks in the evaluated VR platform.

AI USE STATEMENT

In preparing this submission, we used generative AI tools for LaTeX template migration, targeted language revision, literature lookup, code-to-method documentation, formatting, and figure layout. We did not use generative AI tools to generate new participant data or model-training results for this revision. We reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All studies received institutional ethics approval, and all participants provided informed consent, met the vision and VR-safety criteria, and received fixed compensation. Identifying committee and protocol information is withheld during double-blind review. Because gaze and motion trajectories may enable behavioral re-identification, the planned data release is pseudonymized rather than claimed anonymous and should use controlled access. The dataset supports interaction-target prediction and must not be presented as a measure of mental state or private intent.

REPRODUCIBILITY STATEMENT

Appendix C and Table S1 document data acquisition, quality control, and participant/event accounting. Appendix D specifies causal features and excluded inputs. Appendix E and Table S3 give model and training configurations; Appendix J.1 defines the exact-capacity variants. Appendices G–I describe the frozen predictive, transfer, and counterfactual-policy protocols. Appendix L details participant-paired bootstrap inference, sign-flip tests, and Holm families. Appendix E.3 specifies the frozen policy gates, and Table S20 reports runtime settings. These sections distinguish predictive evaluation, robustness tests, and counterfactual policy estimates. The code and a partial processed-data release have been uploaded to the submission system; the data release is partial because of the supplementary-file size limit.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I. Jordan. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021.

Samantha Aziz, Dillon J. Lohr, Lee Friedman, and Oleg Komogortsev. Evaluation of eye tracking signal quality for virtual reality applications: A case study in the meta quest pro. In *Proceedings of the 2024 Symposium on Eye Tracking Research and Applications*, pp. 1–8, 2024. doi: 10.1145/3649902.3653347.

Efe Bozkir, Süleyman Özdel, Mengdi Wang, Brendan David-John, Hong Gao, Kevin Butler, Eakta Jain, and Enkelejda Kasneci. Eye-tracked virtual reality: A comprehensive survey on methods and privacy challenges. *arXiv preprint arXiv:2305.14080*, 2023. doi: 10.48550/arXiv.2305.14080.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1724–1734, 2014. doi: 10.3115/v1/D14-1179.

Brendan David-John, Candace E. Peacock, Ting Zhang, T. Scott Murdison, Hrvoje Benko, and Tanya R. Jonker. Towards gaze-based prediction of the intent to interact in virtual reality. In *Proceedings of the 2021 Symposium on Eye Tracking Research and Applications*, 2021. doi: 10.1145/3448018.3458008.

Robin Dunn, Larry Wasserman, and Aaditya Ramdas. Distribution-free prediction sets for two-layer hierarchical models. *Journal of the American Statistical Association*, 118(544):2491–2502, 2023. doi: 10.1080/01621459.2022.2060112.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. *Proceedings of Machine Learning Research*, 70:1321–1330, 2017.

Baosheng James Hou, Lucy Abramyan, Prasanthi Gurumurthy, Haley Adams, Ivana Tosic Rodgers, Eric J. Gonzalez, Khushman Patel, Andrea Colaço, Ken Pfeuffer, Hans Gellersen, Karan Ahuja, and Mar Gonzalez-Franco. Online-eye: Multimodal implicit eye tracking calibration for xr. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2025. doi: 10.1145/3706598.3713461.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753. PMLR, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Anish S. Narkar, Jan J. Michalak, Candace E. Peacock, and Brendan David-John. Gazeintent: Adapting dwell-time selection in vr interaction with real-time intent modeling. *Proceedings of the ACM on Human-Computer Interaction*, 8(ETRA), 2024. doi: 10.1145/3655600.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3581–3591, 2020.

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*, volume 30, 2017.

Yushi Wei, Rongkai Shi, Difeng Yu, Yihong Wang, Yue Li, Lingyun Yu, and Hai-Ning Liang. Predicting gaze-based target selection in augmented reality headsets based on eye and head endpoint distributions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2023. doi: 10.1145/3544548.3581042.

Zhenyu Yu, Chunlei Meng, Yangchen Zeng, Mohd Yamani Idna Idris, Jihong Guan, and Shuigeng Zhou. CaST-POI: Candidate-conditioned spatiotemporal modeling for next POI recommendation. *arXiv preprint arXiv:2604.20845*, 2026. doi: 10.48550/arXiv.2604.20845.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1059–1068. ACM, 2018. doi: 10.1145/3219819.3219823.

Guorui Zhou, Na Mou, Ying Fan, Qi Pi, Weijie Bian, Chang Zhou, Xiaoqiang Zhu, and Kun Gai. Deep interest evolution network for click-through rate prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 5941–5948, 2019. doi: 10.1609/aaai.v33i01.33015941.

APPENDIX

A RELATED WORK

VR intention and calibration. Prior work predicts interaction from gaze dynamics, adapts dwell selection from online intent, or models eye–head endpoints (David-John et al., 2021; Narkar et al., 2024; Wei et al., 2023). Fit-sensitive eye tracking also motivates implicit online correction (Aziz et al., 2024; Hou et al., 2025). CALM-VR instead ranks the current legal 3D object set from causal multimodal history; its bounded two-angle update is a deployment adjunct.

Target-conditioned history modeling. DIN uses a candidate-dependent activation mechanism to aggregate behavioral history (Zhou et al., 2018), and DIEN models target-relative interest evolution through attention-modulated recurrence (Zhou et al., 2019). The recent CaST-POI preprint also uses candidate-conditioned history readout with candidate-relative geographic bias (Yu et al., 2026). CALM-VR follows the broader target-conditioned modeling principle, but evaluates candidate-specific readout after modality-specific encoding of continuous sensor histories. Our broadcast-query control retains individual candidate tokens in the scorer, while the crossed-summary analysis separately varies candidate alignment in the input summaries. These experiments characterize an architectural choice within CALM-VR; they do not establish superiority over DIN or DIEN, and the broadcast-query control is not an implementation of either model.

Selective and set-structured prediction. Selective and conformal prediction trade action rate for error risk (Geifman & El-Yaniv, 2017; Romano et al., 2020; Angelopoulos et al., 2021). Deep Sets and Set Transformer model unordered inputs (Zaheer et al., 2017; Lee et al., 2019); set structure and temporal readout are distinct modeling choices. CALM-VR tests candidate-specific retrieval from GRU-encoded histories against a parameter-matched shared readout. Both retain candidate-conditioned final scoring, and the crossed-summary analysis distinguishes the readout choice from candidate alignment in the input features.

B SCOPE AND EVIDENCE STAGES

This supplement separates three analysis stages. First, data from 97 participants were used for network fitting, and participant-grouped five-fold out-of-fold (OOF) results supported model/epoch selection. Second, a 21-participant set supported post-hoc development diagnostics and the disjoint temperature/APS fits. Third, one system was frozen before any author accessed data from 30 previously unseen participants; only then were those records opened for offline predictive evaluation and counterfactual policy analysis. The legacy 21-participant partition had already been opened in an earlier pipeline and was excluded from fitting, selection, calibration, predictive evaluation, and inference.

The original data collection was completed by 150 participants. Analytic quality control retained 139 sessions: 97 training/CV, 21 validation, and 21 legacy holdout, with 11 exclusions. The pseudonymized original-cohort dataset contains 803,042 labeled time samples, 66,442 candidate rows, and 5,282 protocol-defined target-bearing selections (139×38).

The independently acquired 30-person cohort was first accessed after freezing the model, calibration, and predictive protocol. Sections G, G.1, H, I, and J report frozen-model offline prediction and counterfactual-policy results. The parameter-matched full-versus-broadcast comparison is a prespecified secondary architectural analysis. Task-balanced, candidate-set, and zero-shot transfer comparisons are secondary analyses; the 97-person OOF and 21-person validation comparisons provide development evidence.

Throughout this supplement, "full query" denotes the candidate-specific member of an exact parameter-matched pair; "complete-system contrast" denotes CALM-VR versus the designated late-fusion Transformer–GHC configuration.

Table S1: Dataset accounting and cache summary. Event counts are from the frozen 1500-ms history cache.

Partition/status	Participants	500-ms events	800-ms events	1000-ms events	Role	Reported use
Training manifest	97	3,686	3,667	3,580	grouped 5-fold CV	selection
Validation manifest	21	798	791	778	post-hoc five-seed	development
Legacy holdout	21	–	–	–	previously opened	not evaluated
QC exclusions	11	–	–	–	failed session QC	absent
Previously unseen evaluation	30	1,131	1,123	1,114	first opened after freeze	offline evaluation

C Acquisition System and Tasks

C.1 Runtime Architecture

The PC-based Unity application runs a Meta Quest Pro with left and right Touch Pro controllers. Persistent services manage participant/session identifiers, condition scheduling, safety state, object registration, timestamps, logs, and prediction requests. The implemented scenes are `S00_Boot`, `S01_SignalCheck`, `S10_TargetAcquisition`, `S20_ClutterShelf`, and `S30_SequentialAssembly`. A shared monotonic clock links frame records, interaction events, candidate snapshots, prediction anchors, responses, and applied actions.

The collection stores abstract gaze origin/direction and quality, HMD/controller poses and motion, validity masks, candidate geometry and state, scene seeds, and Unity interaction events. It excludes eye images, passthrough video, audio, participant names, and device serial numbers.

C.2 Tasks and Quality Gates

T0 validates the sensing stack with nine fixation points, HMD stability, and controller integrity. The analytic protocol requires at least 80% gaze validity and mean fused angular error below 2°. T1 presents ten target-acquisition trials with 12 candidates. T2 presents ten shelf-search trials with 25 candidates, depth variation, partial occlusion, and similar distractors. T3 contains three six-step assembly sequences with six legal candidates. The protocol therefore specifies 38 target-bearing selections (10+10+18). The 500-ms caches retain 38 targets per modeled original-cohort participant; 23 of the 139 analytic raw logs contain 39–40 valid selection records because retry or recovery records are preserved for descriptive QC.

Prediction anchors must precede the selection event by their nominal horizon within a 20-ms tolerance. Events lacking a valid anchor or target, or containing an invalid candidate set, are excluded by the cache builder. Candidate overflow did not occur. Participant, session, trial, sequence, scene label, input device, selection timestamp, anchor timestamp, and realized horizon are explicitly forbidden model inputs.

The freeze sequence was: (1) 97-person grouped CV selected the architecture and fixed six-epoch rule; (2) temperature and APS were fit on disjoint subsets of the 21-person development partition; (3) the evaluation system and checkpoint rule, model and baseline definitions, features, hyperparameters, budgets, endpoints, comparators, participant-equal metrics, CIs/tests, and five-comparison complete-system Holm family were locked, including the full and broadcast-query variants, their training and checkpoint-selection rules, all three horizon contrasts, and the separate Holm–3 query family; and (4) only then was the first 30-person file opened. CALM-VR versus designated late fusion at 500 ms remained the primary predictive comparison; the query comparison was prespecified as secondary. The evaluation retained 1,131/1,123/1,114 of the planned 1,140 events at 500/800/1000 ms; decreasing retention reflects horizon-specific causal-history coverage and protocol exclusions, with no label imputation. The 500-ms task counts are reported with the task-balanced results below.

Table S2: Frozen model inputs. The 49-D candidate tensor uses HMD-relative features and excludes raw world centers.

Group	Dim.	Representative fields	Frame/handling
Gaze history	13	direction, delta, origin, quality, valid, time	HMD-relative; masked
HMD history	12	velocity, angular velocity, forward, valid, time	motion in HMD frame
Each controller	15	position, forward, velocity, angular velocity, valid, time	HMD-relative; masked
Candidate base	13	relative position/direction, distance, extent, visible/occluded/interactable	shared across objects
Candidate alignment	36	six summaries $\times$ head/gaze/L/R directions and L/R velocity	causal history only
Candidate world center	3	x, y, z	stored, excluded from model

D Feature Construction

Each event uses a 1500-ms causal window with 91 temporal samples (approximately 60 Hz) and a maximum of 25 candidates. The gaze stream has 13 features: HMD-frame direction, direction change, HMD-frame origin, gaze quality, gaze validity, time offset, and history validity. The HMD stream has 12 features: linear and angular motion in the anchor-HMD frame, anchor-HMD-frame forward direction, validity, time offset, and history validity. Each controller stream has 15 features: HMD-frame position, forward direction, linear and angular motion, validity, time offset, and history validity.

The raw cache records 52 candidate features. The frozen relative-coordinate configuration removes three world-center coordinates, yielding a 49-dimensional candidate input. It retains relative position/direction, log distance, log extent, visibility, occlusion, interactability, and six summary statistics for each of head direction, gaze, left/right controller direction, and left/right controller velocity: last/mean/max cosine, cosine trend, last perpendicular distance, and valid fraction. Continuous transforms are computed from causal history; absent candidates are padded and masked before loss and evaluation.

E Model and Optimization Details

E.1 Candidate-Query Network

A shared $49 \to 192 \to 192$ candidate MLP creates one token per valid object. Four separate two-layer GRUs encode gaze, HMD, left-controller, and right-controller histories. Each candidate token queries each temporal stream with four-head multi-head attention. The reliability router receives, for each stream, the full-window valid fraction, latest-sample validity, valid fraction over the latest six cached samples, and the horizon embedding; its event-level softmax weights combine the four object-conditioned vectors. The scorer concatenates candidate token, routed evidence, and horizon representation to produce one masked logit per object. There is no endpoint-regression head and no semantic label input in the frozen reported model.

Proof of Proposition 1. For global rotation $A \in SO(3)$ and translation b, let $p'_j = Ap_j + b$, $p'_H = Ap_H + b$, and $R'_H = AR_H$. Then $(R'_H)^\top(p'_j - p'_H) = R_H^\top A^\top A(p_j - p_H) = r_j$. For every jointly transformed direction $v' = Av$, $(R'_H)^\top v' = R_H^\top v$, and $(Av)^\top(Aw) = v^\top w$; hence the stated relative directional and inner-product features are preserved. For ordering, the shared encoder acts row-wise, so $\phi(P\mathbf{C}_{\mathrm{obj}}) = P\phi(\mathbf{C}_{\mathrm{obj}})$. Cross-attention treats candidates as independent queries, hence $\mathrm{MHA}(Pq, Z, Z) = P\,\mathrm{MHA}(q, Z, Z)$. The router is independent of candidate order, the scorer is shared row-wise, and masked softmax in Eq. (2) commutes with a joint permutation. Thus each physical object retains its score and probability. The Top-1 object is also unchanged when the maximizer is unique, or when tie-breaking is permutation-equivariant.

Table S3: Frozen CALM-VR and baseline configurations. Neural rankers share the optimization settings described below.

Setting	Value
Model	CALM-VR (candidate-specific temporal readout)
Coordinate frame	HMD-relative
History / steps	1500 ms / 91
Hidden width / attention heads	192 / 4
Temporal encoders	four separate 2-layer GRUs
Candidate input / maximum	49-D / 25
Hidden-layer dropout / modality dropout	0.1 / 0.0
Semantic features	disabled
Reference implementation size	2,583,749 parameters / 9.88 MiB
Network-fitting participants	97
Frozen evaluation unit	one system locked before cohort access
Fixed final epochs	6

Baseline	Structure and configuration
Gaze ray	Deterministic selection of the legal object with minimum angular deviation from the valid fused gaze ray. No trainable parameters.
H+C	HMD and left/right controller streams, each with a 2-layer GRU of hidden width 192. Candidate-conditioned masked scorer, dropout .1, 1.986M parameters.
LightGBM–GHC	LambdaRank, 500 trees, learning rate .05, 31 leaves, minimum leaf size 20, feature fraction .9, bagging fraction .9.
GRU–GHC	2-layer GRU, hidden width 192. Candidate-independent temporal encoding followed by a candidate-conditioned scorer.
TCN–GHC	Four causal residual blocks, width 192, kernel size 3, dilations 1/2/4/8, dropout .1.
Late-fusion Transformer–GHC	Two Transformer layers, width 192, four attention heads, feed-forward width 384, dropout .1. Modality histories are fused before candidate-conditioned final scoring.
MoE–GHC	Four modality experts of width 192 and a softmax validity/horizon gate. The fused history vector is shared across candidates.

E.2 Loss and Training

The objective is masked candidate cross-entropy plus Brier loss with coefficient 0.1. The frozen run sets consistency loss and expert auxiliary loss to zero. Primary selections receive weight 2.0 and auxiliary selections 0.5. CALM-VR and the neural baselines use AdamW with learning rate 7×10^{-4}, weight decay 10^{-4}, batch size 256, six epochs, cosine decay, gradient clipping at 5, and FP32 arithmetic. Architecture and hyperparameter selection use only participant-grouped CV on the 97-person fitting partition. The evaluation system and checkpoint rule were frozen before access to the 30-person cohort. Seeds 20260720–20260724 support five-seed post-hoc development analysis. The seed-20260720 reference benchmark completed six epochs in 59.57 seconds on an NVIDIA RTX 4080 SUPER using PyTorch 2.10.0 with CUDA 12.8.

Proposition 3 (strict propriety). Fix an input–horizon pair with a nonempty legal set, and let π be its true conditional target distribution. Over all probability vectors p on that legal set, the excess conditional risk is

$$\mathbb{E}_\pi \ell_\lambda(p, Y) - \mathbb{E}_\pi \ell_\lambda(\pi, Y) = D_{\mathrm{KL}}(\pi \| p) + \lambda \|p - \pi\|_2^2 \geq 0. \tag{S1}$$

Proof. Cross-entropy equals $H(\pi) + D_{\mathrm{KL}}(\pi \| p)$. Expanding the expected Brier term gives $1 - \|\pi\|_2^2 + \|p - \pi\|_2^2$. Both excess terms vanish jointly only at $p = \pi$, so π is the unique Bayes minimizer. Positive horizon weights preserve each horizon's Bayes target in the unrestricted function class. Empirical probability quality is evaluated in Appendix F.3.

E.3 Post-Selection Bias Updates and Selective Interface

The controlled synthetic-drift replay uses the lightweight online 2D rule reported in the main paper. After completed selection k, it clips the direct yaw–pitch discrepancy between the selected-object direction and observed gaze to $[-3.5^\circ, 3.5^\circ]^2$, then applies $\delta_k = .6\delta_{k-1} + .4e_k$. Because the implementation takes this difference without angular wrapping, the analysis assumes yaw values do not straddle the `atan2` branch cut. The new state is available only to later trials. This clipped EMA generates the synthetic-drift and adaptation-history results in Tables S18 and S19 and the physical re-donning result reported below; the counterfactual policy replay uses the optimizer below.

Angle map, application, and boundedness. Let $a(v) = (\mathrm{atan2}(v_x, v_z), \mathrm{atan2}(v_y, \sqrt{v_x^2 + v_z^2}))$ map an HMD-frame unit vector to yaw–pitch angles. After completed selection k, the observed gaze g_k and selected-object direction d_{y_k} define the update in Eq. (7), with $\gamma = .6$, $b = 3.5^\circ$, and $\delta_0 = 0$ in the controlled drift replay. The replay implements the angular difference literally, so the analysis assumes yaw values do not straddle the `atan2` branch cut; a deployment spanning that cut must wrap yaw before clipping. The state δ_k becomes available only after selection k and is used from trial $k + 1$ onward; a prediction within trial k adds δ_{k-1} in yaw–pitch space and reconstructs the unit direction. The replay applies this correction to every valid gaze-history vector and recomputes the associated candidate gaze-alignment fields. Network weights remain frozen.

Proposition 2 (bounded causal correction). Given $\delta_0 = 0$ and $0 \le \gamma < 1$, Eq. (7) satisfies $\|\delta_k\|_\infty \le b$. If $e_k = e$ is constant, then $\delta_k = (1 - \gamma^k)e$, so the bias approaches e geometrically. Predictions at trial k depend only on selections completed before that trial.

Proof. Unrolling gives $\delta_k = (1 - \gamma) \sum_{i=1}^{k} \gamma^{k-i} e_i$. Its nonnegative weights sum to $1 - \gamma^k \le 1$, and every e_i lies in the convex box $[-b, b]^2$, proving the bound. Substituting constant $e_i = e$ evaluates the geometric series. Because e_i is formed only after selection i completes, δ_{k-1} is measurable at every anchor in trial k and cannot contain the current-trial target Y_k.

The counterfactual Unity-policy adapter also estimates only yaw and pitch and uses the available prefix of at most the first ten valid completed selections. Each proposed update takes three projected Adam steps at learning rate 0.005 rad with $\lambda = .01$ and clips each angle to $[-3^\circ, 3^\circ]$. An update is retained only if the available-prefix objective decreases by at least 1%; NaN/Inf, invalid tracking, an illegal candidate, or a bound violation causes immediate rollback. Updates are attempted only through the tenth valid completed selection; the accepted bias is then frozen. Network parameters remain fixed, state is initialized identically across policy replays, and updates never use the current trial target before selection. Median/P95 update time is 0.10/0.16 ms. Projection and rollback bound the state and enforce improvement on the current buffer, but do not imply convergence across changing buffers.

The counterfactual Unity-policy replay used a distinct bounded optimizer. Let $B = [-3^\circ, 3^\circ]^2$ and L_k be the loss on the available prefix of at most ten completed selections. Starting from $\delta^{(0)} = \delta_{k-1}$, three Adam steps have the projected form

$$\delta^{(r+1)} = \Pi_B\left(\delta^{(r)} - \eta \frac{\widehat{m}_r}{\sqrt{\widehat{v}_r} + \epsilon}\right), \qquad r = 0, 1, 2. \tag{S2}$$

The proposal $\widetilde{\delta}_k = \delta^{(3)}$ is accepted only if $L_k(\widetilde{\delta}_k) \le .99 L_k(\delta_{k-1})$; otherwise $\delta_k = \delta_{k-1}$. Projection and rollback imply $\delta_k \in B$ by induction. An accepted step improves the *current* completed-selection buffer, but changing buffers precludes a global convergence claim. Updates were attempted only through the first ten valid completed selections in each policy replay; the accepted bias was then frozen. Bias and history were initialized identically across replays; full constants appear in the supplement. Synthetic drift, history-count, and re-donning results use Eq. (7), whereas the counterfactual policy replay uses Eq. (S2).

The frozen selective policy uses APS rather than RAPS because RAPS produced no singleton at the evaluated $\alpha \in \{.05, .10, .20\}$. Temperature was fit on 380 events from 10 development participants and one pooled APS quantile for the frozen offline policy was fit on all 418 events from 11 disjoint development participants; neither subset overlaps the 30-person evaluation cohort. Participant separation prevents subject overlap but does not establish pooled event-level calibration–test exchangeability because repeated events are clustered within participant and ordered across T1–T3.

Table S4: Five-fold OOF development results used for selection.

Endpoint	CALM-VR	H+C	Δ	95% CI
500 ms	53.85	33.72	+20.13	[18.24,22.04]
800 ms	20.62	14.05	+6.57	[5.08,8.04]
1000 ms	20.15	13.28	+6.87	[5.40,8.36]
Equal-horizon	31.54	20.35	+11.19	[10.16,12.23]

The pooled event quantile is therefore used as a frozen empirical decision rule; no finite-sample conformal coverage guarantee is claimed for this repeated-measures design. The final operating point is APS $\alpha = .20$ and triggers only for a singleton candidate stable for $K = 3$ consecutive frames, at least 80% gaze validity over the latest ten samples, valid HMD/active-controller tracking with sample age at most 50 ms, model latency at most 16.67 ms, model-to-Unity latency at most 33.3 ms, and a final legal-candidate check. Failure of any gate causes abstention.

APS construction and conditional guarantee. The frozen action policy fits temperature T on 380 events from 10 development participants and one pooled adaptive-prediction-set (APS) quantile (Romano et al., 2020) on $n = 418$ events from 11 other development participants. Write $\bar{p}_T = \mathrm{softmax}(z/T)$. Scalar temperature scaling with $T > 0$ preserves $\arg\max_{j \in \mathcal{O}_t} z_j$ for fixed logits and legal support. CALM-VR therefore has identical 500/800/1000-ms Top-1 before and after scaling: 51.20/19.00/19.10%, with a 0.00-pp change at each horizon (Table 2). Calibration changes probabilities and the set-valued interface, not Top-1 decisions. For calibration event i, let $j_{i,(1)}, \ldots, j_{i,(N_i)}$ order candidates by decreasing $\bar{p}_{T,i}$, and let $Y_i = j_{i,(r_i)}$. Define

$$S_i = \sum_{\ell \le r_i} \bar{p}_{T,i}(j_{i,(\ell)}), \qquad \hat{q} = S_{(\lceil (n+1)(1-\alpha) \rceil)}. \tag{S3}$$

For a test input whose candidates are ordered as $j_{(1)}, \ldots, j_{(N)}$, use

$$\Gamma_\alpha(x) = \left\{ j_{(r)} : \sum_{\ell < r} \bar{p}_T(j_{(\ell)}) < \hat{q} \right\}, \tag{S4}$$

Proposition 4 (conditional event-level APS coverage). Condition on a predictor and temperature fixed independently of the APS-calibration labels. If the n calibration event pairs (X_i, Y_i) and one future event pair are exchangeable, then $\Pr\{Y \in \Gamma_\alpha(X)\} \ge 1 - \alpha$. Any policy that acts only by selecting the sole member of $\Gamma_\alpha(X)$ therefore satisfies

$$\Pr\{\text{wrong action and } |\Gamma_\alpha(X)| = 1\} \le \alpha. \tag{S5}$$

Proof. Exchangeability and the split-conformal rank lemma give $\Pr\{S_{n+1} \le \hat{q}\} \ge 1 - \alpha$; ties are conservative. For positive softmax scores, this event implies $Y \in \Gamma_\alpha(X)$. A wrong singleton action implies a set miss, proving both claims.

Here $S_{(k)}$ is the kth order statistic. Equation (S5) is only an unconditional per-event bound, not trigger-conditional or session-wise. Participant separation does not establish exchangeability for clustered, T1–T3-ordered trials (Dunn et al., 2023); hence the current APS rule is empirical, with no claimed finite-sample validity. We instead report participant-level policy risks. The evaluated $\alpha = .20$ policy also applies the frozen gates in Figure 2.

F Development Evidence

F.1 Participant-Grouped Cross-Validation

The selection metric gives equal weight to each participant and to the 500/800/1000-ms horizons. Events are paired within fold, and intervals cluster by participant with 50,000 bootstrap draws. These grouped-CV summaries characterize model-selection performance.

Table S5: Five-seed post-hoc validation (Top-1, %).

Endpoint	CALM-VR	H+C	Δ	Participant CI
500 ms	52.48	32.11	+20.37	[17.31,23.62]
800 ms	19.38	12.95	+6.43	[4.52,8.44]
1000 ms	19.44	12.23	+7.21	[4.79,9.68]
Equal-horizon	30.43	19.10	+11.33	[9.40,13.43]

Table S6: Comparisons in the frozen offline evaluation of 30 previously unseen participants.

Endpoint	Comparison	Δ	95% CI	Holm p	Decision
500 ms	CALM-VR − designated late-fusion Transformer–GHC	+4.78	[2.88,6.68]	$< .001$	significant
500 ms	CALM-VR − H+C	+19.60	[16.50,22.70]	$< .001$	significant
500 ms	CALM-VR − gaze ray	+11.60	[9.10,14.10]	$< .001$	significant
800 ms	CALM-VR − designated late-fusion Transformer–GHC	+5.60	[4.20,7.00]	$< .001$	significant
1000 ms	CALM-VR − designated late-fusion Transformer–GHC	+5.90	[4.40,7.40]	$< .001$	significant

F.2 Post-Hoc Validation

The five frozen seeds were evaluated on 21 validation participants. The conditional interval resamples participants while holding seeds fixed; a joint interval resamples both axes. This previously inspected partition provides post-hoc evidence of development stability.

The corresponding joint seed-by-participant CIs are $[16.83, 24.05]$, $[4.20, 8.75]$, $[4.52, 9.92]$, and $[9.20, 13.59]$ pp; all exclude zero. These comparisons support development stability; the incremental value of candidate-specific readout is assessed by the parameter-matched controls in Section J.

F.3 Probability Quality and Selective Behavior

CALM-VR improved predictive and probability-quality metrics over H+C in the 21-person post-hoc partition. Equal-horizon averages over 500, 800, and 1000 ms were 30.43% versus 19.10% Top-1, 56.97% versus 46.29% Top-3, 31.51% versus 16.74% Macro-F1, 32.07% versus 16.94% balanced accuracy, and 3.96% versus 8.65% ECE. The five-seed participant-equal Top-1 values were 52.48%, 19.38%, and 19.44% across horizons. Matched summaries were descriptively close between development and frozen evaluation: 52.48% versus 51.20% at 500 ms and 30.43% versus 29.77% with equal horizon weighting (a 0.66-pp difference).

On the disjoint APS split, confidence ranking yielded the lowest selective risk at matched 5/10/20% action rates: 0.0/1.7/2.7% (AURC .207), compared with 0.0/2.5/4.8% for raw APS (.218) and 0.3/3.0/5.6% for scaled APS (.226). The corresponding confidence wrong-operation rates were 0.00/0.17/0.54%, unchanged by temperature scaling. Confidence provides the risk-ranking reference, while APS defines the frozen empirical set-valued interface. The counterfactual policy evaluates this interface with its additional action gates at the 6.5% operating rate.

G Frozen 30-Participant Offline Predictive Evaluation

The offline comparison includes gaze-ray, H+C, the designated late-fusion Transformer–GHC comparator, and CALM-VR. The 500-ms comparison against this designated comparator is the principal endpoint. The 800- and 1000-ms complete-system intervals also exclude zero. Prespecified secondary parameter-matched comparisons show smaller positive query gains at these horizons (Table S14); these query increments are distinct from the complete-system gains. The designated late-fusion comparator is fixed across horizons.

Table S7: Task-specific, task-macro, and ranking results in the frozen 30-person evaluation at 500 ms. Δ_{LF} compares full query with late fusion; Δ_Q compares full with broadcast query. Accuracy is computed within participant and task before task averaging. Displayed CIs compare full query with late fusion; query task contrasts are descriptive point estimates. MRR and nMRR are descriptive.

Task / nominal K	Events	Chance (%)	CALM-VR Top-1 (%)	Late fusion Top-1 (%)	Δ_{LF} (pp) 95% CI
T1 target acquisition / 12	298	8.33	48.00	44.00	+4.00 [1.00,7.00]
T2 cluttered-shelf search / 25	297	4.00	56.00	50.00	+6.00 [2.40,9.60]
T3 sequential assembly / 6	536	16.67	50.32	45.78	+4.54 [1.90,7.18]
Three-task macro	1,131	9.67	51.44	46.59	+4.85 [2.70,7.00]
Event-weighted overall	1,131	11.14	51.20	46.42	+4.78 [2.88,6.68]

Task / nominal K	Broadcast Top-1 (%)	Δ_Q (pp)	CALM-VR MRR	Late fusion MRR	CALM-VR nMRR	Late fusion nMRR
T1 target acquisition / 12	51.56	-3.56	.680	.650	.568	.528
T2 cluttered-shelf search / 25	54.40	+1.60	.720	.680	.670	.622
T3 sequential assembly / 6	44.17	+6.15	.700	.660	.493	.425
Three-task macro	50.04	+1.40	.700	.663	.577	.525
Event-weighted overall	48.80	+2.40	.700	.663	.559	.504

G.1 Task-Balanced and Candidate-Set Evaluation

As a secondary analysis of the frozen 30-person evaluation, CALM-VR exceeded the designated late-fusion comparator in T1, T2, and T3 at 500 ms. Equal task weighting yielded 51.44% versus 46.59%, close to the event-weighted 51.20/46.42% result. Retained events preserve the nominal legal-set size within each task, so task-row chance is $1/K$ and the overall chance value averages event-specific $1/K$. For legal-set size K, chance-normalized reciprocal rank is

$$\mathrm{nMRR} = \frac{\mathrm{MRR} - H_K/K}{1 - H_K/K},$$

where H_K is the Kth harmonic number; normalization precedes participant and task aggregation.

Candidate-specific readout has a positive point estimate under equal task weighting: full and broadcast query average 51.44% and 50.04%, respectively. The task-specific contrasts are -3.56 pp in T1, $+1.60$ pp in T2, and $+6.15$ pp in T3, varying in direction and magnitude across the observed tasks. These query contrasts are descriptive point estimates, distinct from the full-versus-late-fusion intervals in Table S7. The $+1.40$ pp task-macro summary weights tasks equally, whereas the $+2.40$ pp overall contrast (51.20% versus 48.80%) weights retained events within each participant before averaging participants. The protocol schedules 10:10:18 selections across T1–T3, with 298/297/536 retained events. Using the planned composition gives the arithmetic approximation $[10(-3.56) + 10(1.60) + 18(6.15)]/38 \approx 2.40$ pp, while equal task weighting gives $(-3.56 + 1.60 + 6.15)/3 \approx 1.40$ pp. Reported metrics use the actual retained events and participant-level aggregation. The largest point-estimated query benefit occurs in six-candidate assembly, not 25-candidate search; this task contrast does not isolate the contribution of candidate count or other task properties.

Candidate-set sensitivity protocol. The sweep reuses the 297 retained T2 cluttered-shelf events at the 500-ms horizon from the frozen 30-participant evaluation. The target event, causal sensor history, model parameters, and target identity remain fixed across $K \in \{2, 4, 6, 8, 12, 16, 25\}$. Distractors come from the non-target objects in the event's original anchor-time T2 legal set. Let r_j denote a candidate's HMD-relative position and $\widehat{r}_j = r_j/\|r_j\|_2$. Relative to target y, distractors are ordered by increasing angular separation, with absolute log-distance difference breaking angular ties:

$$\theta_j = \arccos(\widehat{r}_j^{\top}\widehat{r}_y), \qquad \ell_j = \left|\log\|r_j\|_2 - \log\|r_y\|_2\right|.$$

The target and the first $K-1$ distractors under this lexicographic order form each nested candidate set. The ordering uses anchor-time geometry rather than model scores or learned difficulty estimates. Thus $K = 25$ recovers the original T2 condition, and every setting uses the same 297 target events. Target identity defines the controlled candidate subsets and remains excluded from predictor inputs.

Table S8: Descriptive nested candidate-set sweep on the same 297 T2 events at 500 ms in 30 participants. Each subset contains the target and its $K-1$ geometrically nearest distractors. The $K=25$ row reproduces the original T2 condition in Table S7.

Legal candidates K	Random Top-1 (%)	Full query (%)	Broadcast query (%)	Query gain (pp)	Late fusion (%)
2	50.00	76.0	75.7	+0.3	73.5
4	25.00	67.5	67.0	+0.5	63.5
6	16.67	63.5	62.8	+0.7	59.0
8	12.50	61.5	60.6	+0.9	56.5
12	8.33	59.0	57.9	+1.1	53.5
16	6.25	57.5	56.2	+1.3	51.5
25	4.00	56.0	54.4	+1.6	50.0

The full-minus-broadcast point estimate rose from $0.3\,\mathrm{pp}$ at $K=2$ to $1.6\,\mathrm{pp}$ at $K=25$, with all accuracies above uniform chance. The fixed-event comparison isolates sensitivity to candidate-list composition as additional legal alternatives enter the nested sets.

H Zero-Shot Task and Layout Generalization

Zero-shot transfer protocol. Transfer trials were collected from the same 30 model-unseen participants as the frozen evaluation, separately from the standard T1–T3 prediction cache. The held-out-layout condition preserved seen interaction logic while changing scene layouts and object configurations. It planned 38 selections per participant and retained 1,128 target events at 500 ms under the same anchor-time legality and causal-history QC. U1 was a held-out single-stage selection flow with $K=12$; U2 was a held-out multi-stage, state-dependent selection flow with $K=18$. Each task planned 20 target-bearing selections per participant, retaining 596 and 594 events, respectively (Table S9). All transfer trials were excluded from network fitting, calibration, baseline and hyperparameter selection, checkpoint selection, and policy selection. Each frozen model was evaluated without retraining or online adaptation.

Accuracy is computed per participant before population averaging, and the unseen-task macro weights U1 and U2 equally. Chance accuracy is computed as $1/K$ from the realized legal-set size before aggregation. The U1/U2 task-macro chance is therefore $(1/12+1/18)/2=6.94\%$, while the 8.64% held-out-layout value reflects its realized candidate-set composition.

CALM-VR exceeded both the broadcast-query control and designated late fusion in each transfer condition. Across unseen tasks, task-macro Top-1 was 44.65%, 42.75%, and 39.65%, respectively. Query-gain intervals summarize participant-level uncertainty in this secondary analysis; full-versus-late-fusion differences are descriptive point estimates.

The seen-task $+1.40\,\mathrm{pp}$ value gives T1–T3 equal weight, whereas the $+2.40\,\mathrm{pp}$ exact-capacity endpoint weights retained events within participant (Tables S7 and S14). The unseen-task macro uses the same task-equal convention and yields a $+1.90\,\mathrm{pp}$ query gain over the broadcast control.

I Counterfactual Policy Evaluation

The frozen No-assist, Always top-1, and Selective CALM-VR policies were applied offline to the 30-person records. Policy-specific validity filtering retained 1,135, 1,132, and 1,131 opportunities, respectively. All results in this section are counterfactual estimates.

Estimated post-selection wait time spans selection confirmation to resource/interaction readiness. Correct prefetch and waste are conditional on a counterfactual trigger; the action rate is the fraction of valid opportunities on which the policy would trigger. No-assist never triggers, so its hit/waste rates are not applicable and its wrong-trigger and action rates are zero by policy definition. Always top-1 has a 100% counterfactual action rate by definition. Participant-level rates were computed before aggregation.

Table S9: Zero-shot transfer protocol and secondary results at 500 ms. The same 30 evaluation participants completed separate transfer trials. All models remain frozen, without retraining or adaptation. Query intervals are participant-level 95% CIs; OOD comparisons are outside the confirmatory Holm families.

Transfer condition	Participants	Planned selections	Retained events	Legal set
Held-out layouts, seen logic	30	38/person = 1,140	1,128	realized K
U1: single-stage flow	30	20/person = 600	596	$K = 12$
U2: multi-stage flow	30	20/person = 600	594	$K = 18$

Evaluation setting	Chance (%)	Full query (%)	Broadcast query (%)	Late fusion (%)
Seen-task macro (task-equal)	9.67	51.44	50.04	46.59
Unseen layouts, seen task logic	8.64	48.20	46.60	43.30
Unseen task U1 / $K = 12$	8.33	46.50	44.80	41.80
Unseen task U2 / $K = 18$	5.56	42.80	40.70	37.50
Unseen-task macro	6.94	44.65	42.75	39.65

Evaluation setting	Query gain (pp) 95% CI where shown	Full vs. late fusion (pp)
Seen-task macro (task-equal)	+1.40	+4.85
Unseen layouts, seen task logic	+1.60 [0.45,2.75]	+4.90
Unseen task U1 / $K = 12$	+1.70 [0.40,3.00]	+4.70
Unseen task U2 / $K = 18$	+2.10 [0.65,3.55]	+5.30
Unseen-task macro	+1.90 [0.75,3.05]	+5.00

Table S10: Counterfactual policy estimates in 30 participants. Values are participant-equal means; correct and waste percentages are conditional on triggered prefetches.

Condition	Valid trials	Triggers	Wait (ms)	Correct (%)	Waste (%)	Wrong/min	Action rate (%)
No assist	1,135	0	1,418.0	–	–	0 (defined)	0 (defined)
Always top-1	1,132	1,132	1,219.6	51.1	48.9	1.28	100 (defined)
Selective CALM-VR	1,131	74	1,388.5	96.9	3.1	0.01	6.5

Table S11: Participant-paired counterfactual policy contrasts. Difference CIs use 50,000 paired bootstrap draws; two-sided p values use participant-level sign flips.

Family/metric	Comparison	Difference	95% CI	Holm p
A: wait time	Always − No assist	-198.4 ms	[-241.6,-155.2]	$< .001$
A: wait time	Selective − No assist	-29.5 ms	[-50.6,-8.4]	$\approx .006$
A: wait time	Selective − Always	+168.9 ms	[139.9,197.9]	$< .001$
B: correct prefetch	Selective − Always	+45.8 pp	[40.6,51.0]	$< .001$
B: waste	Selective − Always	-45.8 pp	[-51.0,-40.6]	$< .001$
B: wrong triggers	Selective − Always	-1.27/min	[-1.42,-1.13]	$< .001$

The counterfactual policies define different utility–risk operating points. Always top-1 reduced estimated mean waiting by 198.4 ms (14.0%) relative to No-assist, with approximately 554 wasted prefetches among 1,132 triggers. Selective CALM-VR reduced estimated waiting by 29.5 ms (2.1%) while achieving 96.9% correct prefetches, 3.1% waste, and 0.01 incorrect triggers per minute. Relative to Always, this corresponds to 168.9 ms more waiting, $+45.8$ pp correct prefetch, -45.8 pp waste, and -1.27 incorrect triggers/min.

J Ablations and Robustness

J.1 Parameter-Matched Candidate-to-History Assignment

Let $q_j = \phi_\theta(c_j)$ denote the embedded candidate token. For modality m, the candidate-specific model retrieves temporal evidence with

$$z_j^m = \mathrm{MHA}_m(q = q_j, K = H^m, V = H^m). \tag{S6}$$

The exact-capacity broadcast-query (matched-global) control computes the masked mean of the legal embedded candidate tokens,

$$\begin{aligned} \bar{q} &= \frac{\sum_j m_j q_j}{\sum_j m_j}, \\ z_j^m &= \mathrm{MHA}_m(q = \bar{q}, K = H^m, V = H^m), \end{aligned} \tag{S7}$$

and broadcasts this one query to every legal candidate. Importantly, the individual token q_j still enters the final shared scorer. A crossed summary factor either retains the candidate-aligned 36-D history summaries or replaces them with their masked candidate-set mean, broadcast before the same candidate encoder. The cache, legal support, four temporal encoders, all attention blocks, router, scorer, loss, epochs, folds, and training budget are fixed. Consequently every cell has exactly 2,583,749 trainable parameters; the factors change only temporal-query identity and summary alignment.

Table S12: Parameter-matched 2×2 readout analysis (participant-equal Top-1, %). OOF uses 97-person grouped-CV predictions from seed 20260720; validation averages five seeds on 21 participants. Contrasts use unrounded estimates.

Variant	Temporal query	History summaries	Params
Candidate-specific query	candidate-specific	aligned	2,583,749
Broadcast query	shared set mean	aligned	2,583,749
Globalized summaries	candidate-specific	set-mean broadcast	2,583,749
No assignment	shared set mean	set-mean broadcast	2,583,749

Variant	97-person OOF			21-person validation		
	500 ms	800 ms	1000 ms	500 ms	800 ms	1000 ms
Candidate-specific query	53.85	20.62	20.15	52.48	19.38	19.44
Broadcast query	51.44	19.70	19.27	50.13	18.52	18.60
Globalized summaries	52.79	20.42	20.00	51.48	19.23	19.29
No assignment	51.33	19.50	19.12	50.08	18.37	18.45
Query gain: aligned (pp)	+2.41	+0.92	+0.88	+2.36	+0.86	+0.84
Query gain: set mean (pp)	+1.47	+0.92	+0.88	+1.40	+0.86	+0.84
Difference of gains (pp)	+0.95	0.00	0.00	+0.95	0.00	0.00

At 500 ms, all 97 OOF participants contribute 38 events; validation contains $21 \times 38 \times 5 = 3990$ event–seed predictions with equal participant and seed weights. Candidate-specific readout improves Top-1 under either summary representation. With aligned summaries, the query gains are $2.41\,\mathrm{pp}$ in OOF and $2.36\,\mathrm{pp}$ in validation; with set-mean summaries, they are $1.47\,\mathrm{pp}$ and $1.40\,\mathrm{pp}$. The resulting difference of query gains is $0.95\,\mathrm{pp}$ in each partition. This descriptive 2×2 contrast supports complementarity between candidate-aligned summaries and candidate-specific readout at 500 ms.

Table S13: Candidate-specific minus broadcast-query contrasts. CIs cluster by participant. The reported p values are unadjusted two-sided sign-flip tests; the 21-person partition is post-hoc development evidence.

Partition/endpoint	Δ (pp)	Participant 95% CI	Sign-flip p	Interpretation
97 OOF, 500 ms	+2.41	[1.58,3.24]	.0012	strongest development effect
21 validation, 500 ms	+2.36	[1.24,3.46]	.021	concordant post-hoc effect
97 OOF, 800 ms	+0.92	[0.39,1.45]	$\approx .002$	smaller positive effect
21 validation, 800 ms	+0.86	[0.10,1.62]	$\approx .03$	smaller positive post-hoc effect
97 OOF, 1000 ms	+0.88	[0.35,1.41]	$\approx .003$	smaller positive effect
21 validation, 1000 ms	+0.84	[0.10,1.58]	$\approx .03$	smaller positive post-hoc effect

The candidate-query effect was positive across all three horizons in OOF, validation, and the additional 30-person evaluation, with the largest isolated effect at the principal 500-ms endpoint. OOF uses the reference seed, whereas validation averages five fixed seeds. In the 30-person records, candidate-specific and broadcast-query variants achieved 51.20% and 48.80% at 500 ms, respectively ($+2.40$ pp, paired participant-level 95% CI $[1.35, 3.45]$, Holm–3 $p = .027$). At 800 and 1000 ms, full/broadcast accuracies were 19.00/18.10% and 19.10/18.20%, respectively, yielding $+0.90$ pp at both horizons (95% CIs $[0.15, 1.65]$ and $[0.13, 1.67]$; both Holm–3 $p \approx .036$). The development partitions showed concordant smaller effects: $+0.92/+0.88$ pp at 800/1000 ms in OOF and $+0.86/+0.84$ pp in validation. These results support a horizon-dependent rather than horizon-specific contribution of candidate-to-history assignment.

The crossed summary control retains these secondary-horizon query increments under either history-summary representation. In OOF, candidate-specific minus broadcast-query gains are $20.62-19.70 = 0.92$ pp at 800 ms and $20.15 - 19.27 = 0.88$ pp at 1000 ms with aligned summaries; with set-mean summaries, they are $20.42 - 19.50 = 0.92$ pp and $20.00 - 19.12 = 0.88$ pp. The corresponding validation differences are $19.38 - 18.52 = 0.86$ pp and $19.44 - 18.60 = 0.84$ pp with aligned summaries, and $19.23 - 18.37 = 0.86$ pp and $19.29 - 18.45 = 0.84$ pp with set-mean summaries. These descriptive contrasts show that the secondary-horizon query gains persist when candidate-aligned history summaries are replaced by their set mean.

Table S14: Prespecified secondary parameter-matched candidate-query comparison in the frozen 30-person evaluation. Both variants, their training and checkpoint-selection rules, the three contrasts, and the separate Holm–3 family were fixed before cohort access.

Prediction horizon	Full query	Broadcast query	Query gain	Joint participant 95% CI	Holm–3 p
500 ms	51.20	48.80	+2.40	[1.35,3.45]	.027
800 ms	19.00	18.10	+0.90	[0.15,1.65]	.036
1000 ms	19.10	18.20	+0.90	[0.13,1.67]	.036

J.2 Ordered Bridge Attribution

Table S15: Ordered bridge from the designated late-fusion Transformer–GHC baseline to the full CALM-VR configuration at 500 ms. Adjacent effects are conditional on the displayed sequence.

Model configuration	Top-1 (%)	Gain vs. previous (pp)	95% CI
A. Designated late-fusion Transformer–GHC	46.42	Reference	–
B. Capacity- and protocol-matched late fusion	47.62	+1.20	[0.25,2.15]
C. Staged CALM backbone + broadcast query	48.50	+0.88	[0.18,1.94]
D. Staged candidate-specific query	50.90	+2.40	[1.35,3.45]
E. Full CALM-VR configuration	51.20	+0.30	[-0.30,0.90]
E vs. A: complete-system contrast	51.20 vs. 46.42	+4.78	[2.88,6.68]

Along the displayed order, the adjacent differences are $+1.20$, $+0.88$, $+2.40$, and $+0.30$ pp and sum to the complete $+4.78$ pp contrast. Candidate-specific querying is the largest adjacent im-

provement in this path. These differences are conditional on the displayed sequence, rather than unique causal component effects. Bridge stages C/D are intermediate configurations; Table S14 reports the final-protocol candidate-specific/broadcast-query pair. The final step groups the remaining final-configuration changes.

Across horizons, the final-protocol query increments are $+2.40$, $+0.90$, and $+0.90\,\mathrm{pp}$, compared with complete-system gains of $+4.78$, $+5.60$, and $+5.90\,\mathrm{pp}$ at 500, 800, and 1000 ms. Broadcast-minus-late-fusion differences are $+2.38$, $+4.70$, and $+5.00\,\mathrm{pp}$. The query/full-gap ratios are therefore 50.21%, 16.07%, and 15.25%, respectively. These descriptive ratios compare accuracy differences, not unique causal contribution shares. Equal horizon weighting gives 29.77% for full query and 28.37% for broadcast query, a descriptive average query gain of $+1.40\,\mathrm{pp}$.

J.3 Parameter-Matched Set-Aware Baselines

Deep Sets–GHC supplies a low-cost invariant sanity baseline, while Set Transformer–GHC is the external set-structured comparator. In the reported implementations, both consume candidate-independent causal history summaries, emit one dynamic legal-set logit per candidate, and use the same legal mask, loss, event splits, five validation seeds, epoch budget, and candidate features as CALM-VR. Their parameter counts remain within $\pm 5\%$ of the full model.

Table S16: Parameter-matched set-aware comparison at 500 ms (participant-equal Top-1, %). Positive gaps favor full CALM-VR.

Model	Params	97 OOF	21 validation	OOF gap	Validation gap
Full CALM-VR	2,583,749	**53.85**	**52.48**	–	–
Set Transformer–GHC	2,610,432 (+1.03%)	49.58	48.33	+4.27	+4.15
Deep Sets–GHC	2,511,104 (−2.81%)	46.74	45.42	+7.11	+7.06

Every 500-ms set-aware contrast excluded zero: the full-minus-Set Transformer 95% CIs were $[3.38, 5.18]$ (OOF) and $[3.03, 5.25]$ (validation), and the full-minus-Deep Sets intervals were $[6.05, 8.19]$ and $[5.62, 8.48]$, respectively.

Together with the exact-capacity broadcast-query control, these parameter-matched development comparisons support candidate-specific temporal readout beyond model size and unordered-set handling.

J.4 Missing and Corrupted Modalities

Table S17: Frozen CALM-VR 500-ms Top-1 under simulated corruption. Changes are aggregate differences from the clean 51.20% result.

Condition	Top-1 (%)	Change vs. clean (pp)
Clean	51.20	0.00
Missing gaze	32.80	-18.40
300-ms gaze gap	39.80	-11.40
Missing left controller	51.40	+0.20
Missing right controller	49.90	-1.30
Gaze temporal shift	50.80	-0.40

The corruption analysis identifies gaze as the principal information source: removal and a 300-ms gap produce the largest clean-referenced changes ($-18.4\,\mathrm{pp}$ and $-11.4\,\mathrm{pp}$). Controller and temporal-shift perturbations yield smaller aggregate changes. Table S17 characterizes the complete frozen model under each perturbation.

J.5 Synthetic Drift and Adaptation History

Table S18: Synthetic angular drift at 500 ms (Top-1, %).

Drift	None	Online 2D	Small adapter	Oracle
0°	51.20	51.10	50.90	51.10
$\pm1^\circ$	50.20	50.80	50.60	51.10
$\pm2^\circ$	48.40	49.50	49.20	51.00
$\pm3^\circ$	45.80	48.10	47.80	50.90

For Online 2D, the clean, $\pm1^\circ$, $\pm2^\circ$, and $\pm3^\circ$ changes are -0.1, $+0.6$, $+1.1$, and $+2.3$ pp, with 95% CIs $[-.5, .3]$, $[-.1, 1.3]$, $[.2, 2.0]$, and $[1.2, 3.4]$ and adjusted $p = .74$, $.18$, $.036$, and $< .001$, respectively.

For the small adapter, the $\pm2^\circ$ difference is $+0.8$ pp (CI $[-0.2, 1.8]$, $p = .21$) and the $\pm3^\circ$ difference is $+2.0$ pp ($[0.8, 3.2]$, $p = .006$). Oracle correction gives $+2.6$ pp at $\pm2^\circ$ ($[1.7, 3.5]$) and $+5.1$ pp at $\pm3^\circ$ ($[4.0, 6.2]$), both $p < .001$. The online 2D method provides slightly better recovery than the small adapter with far fewer adaptable parameters.

Table S19: Online 2D adaptation versus completed history (Top-1, %).

Selections	Clean	$\pm1^\circ$	$\pm2^\circ$	$\pm3^\circ$
0	51.20	50.20	48.40	45.80
5	51.10	50.60	49.00	47.10
10	51.10	50.80	49.50	48.10
20	51.10	50.90	49.70	48.40
30	51.00	50.90	49.70	48.50

Most recovery occurs within ten completed selections. In the real re-donning condition, unadapted accuracy is 47.7% and the ten-selection online value is 48.6%, a $+0.9$ pp paired difference (95% CI $[0.2, 1.6]$).

Extended drift summary. Synthetic gaze rotation degraded pre-adaptation accuracy monotonically. Under $\pm3^\circ$ drift, accuracy at 0/5/10/20/30 completed selections was 45.8/47.1/48.1/48.4/48.5%. At 0/10/30 selections, it was 51.2/51.1/51.0% without drift, 50.2/50.8/50.9% at $\pm1^\circ$, and 48.4/49.5/49.7% at $\pm2^\circ$. Most recovery occurred by ten selections: at $\pm3^\circ$, the gain was 2.3 pp (95% CI $[1.2, 3.4]$, adjusted $p < .001$), P5 rose from 36.5% to 40.2%, and the worst value from 34.0% to 38.1%, while remaining 2.8 pp below oracle. At $\pm2^\circ$, the online gain was 1.1 pp ($[0.2, 2.0]$, $p = .036$). Physical re-donning yielded a 0.9 pp recovery ($[0.2, 1.6]$).

K Runtime and Reproducibility

Latency uses batch size one, 50 warm-up iterations, and 300 measured iterations on seed 20260720. It excludes Unity serialization and end-to-end scheduling.

Table S20: Model size and single-sample latency (ms). Sizes divide bytes by 2^{20}. GPU: NVIDIA RTX 4080 SUPER; PyTorch 2.10.0/CUDA 12.8.

Model	Params/checkpoint	CPU P50/P95	GPU P50/P95
CALM-VR	2.584M/9.88 MiB	49.14/55.14	12.54/12.62
H+C	1.986M/7.59 MiB	39.34/46.61	9.40/9.42
Multi-scale	5.835M/22.31 MiB	120.20/136.04	23.01/23.22

The evaluated CALM-VR model has 2.584M trainable parameters and a 9.88-MiB serialized checkpoint. On an RTX 4080 SUPER, batch-one GPU P50/P95 latency was 12.54/12.62 ms, versus 9.40/9.42 ms for the 1.986M-parameter H+C model; peak allocated memory was 21.26 and 18.96 MiB, respectively. CALM-VR's CPU P95 was 55.14 ms. These measurements exclude feature

extraction, Unity serialization, communication, scheduling, and actuation; they characterize model inference rather than end-to-end motion-to-action latency.

GPU throughput was 79.68, 106.39, and 43.00 samples/s for CALM-VR, H+C, and Multi-scale, respectively; peak allocated memory was 21.26, 18.96, and 33.87 MiB in the same order.

The implementation uses the fixed model, feature, optimization, and participant-split settings specified in this appendix. The reported development comparisons use five training seeds; evaluation applies the selected frozen configuration. Runtime measurements use the same input shapes and hardware settings across models.

L Statistics, Ethics, and Data Governance

Participant-level inference. For each model contrast and horizon, accuracy is computed separately for every participant. Pairing the two model accuracies gives $d_p = a_p^{(A)} - a_p^{(B)}$ and the population contrast $\bar{d} = P^{-1} \sum_{p=1}^{P} d_p$. For the frozen complete-system and full-versus-broadcast contrasts, 95% percentile confidence intervals use 50,000 participant-paired bootstrap resamples. Each replicate samples P participants with replacement, preserves the within-participant model pairing, and recomputes $\bar{d}$. The 2.5th and 97.5th percentiles form the interval.

Two-sided p values use 50,000 participant-level sign-flip randomization draws. For draw b, independent equiprobable signs $\varepsilon_{bp} \in \{-1, +1\}$ yield $T_b = P^{-1} \sum_p \varepsilon_{bp} d_p$. With $T_{\text{obs}} = \bar{d}$ and $B = 50{,}000$, the finite-simulation correction is

$$p_{\text{raw}} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}[|T_b| \geq |T_{\text{obs}}|]}{B + 1}. \tag{S8}$$

Holm correction is applied within each prespecified hypothesis family. The five complete-system contrasts form one family, and the three full-versus-broadcast horizon contrasts form a separate Holm–3 family. Both families, including their constituent contrasts and statistical procedures, were fixed before the first evaluation-cohort access. CALM-VR versus designated late fusion at 500 ms is the primary predictive comparison; the full-versus-broadcast family is a prespecified secondary architectural analysis. Displayed 95% CIs are percentile intervals for individual contrasts; reported evaluation p values carry the stated family-wise adjustment.

Development and secondary analyses. OOF intervals cluster by participant, and validation summaries average the five fixed seeds before participant-level inference. The 97-person OOF and 21-person validation readout tests use participant-level sign flips and remain separate from the frozen-evaluation families. Validation also reports joint seed-by-participant intervals where indicated. Task-specific and task-macro accuracy contrasts use participant-level intervals where shown. In Table S7, those intervals compare full query with late fusion, not broadcast query; task-level and equal-task full-minus-broadcast contrasts are descriptive point estimates outside the prespecified Holm–3 family. MRR, nMRR, and the candidate-set sweep are descriptive summaries. Zero-shot transfer reports participant-level query-gain intervals outside the confirmatory Holm families. Synthetic-drift adjusted p values use a separate four-comparison family.

Counterfactual-policy difference CIs use 50,000 participant-paired bootstrap draws and two-sided participant-level sign-flip tests. Holm adjustment is separate for the three waiting-time contrasts (family A) and the three Selective-versus-Always risk/efficiency contrasts (family B). These participant-level policy estimates complement the empirical APS interface described in Appendix E.3.

All studies received institutional ethics approval; identifying committee and protocol information is withheld for double-blind review. All participants provided informed consent, had normal or corrected-to-normal vision, reported no condition precluding VR use, and received fixed monetary compensation. The original 150-person cohort comprised 75 women and 75 men; the 30-person evaluation cohort comprised 15 women and 15 men. Across cohorts, age was $M = 25.1$ years (SD $= 5.6$). Withdrawal and exclusion followed the approved protocol, and analytic counts denote completed, included sessions. Because gaze and motion trajectories may enable behavioral re-identification, the planned release is pseudonymized, not anonymous, and should use controlled access. The dataset supports interaction-target prediction only and must not be presented as a measure of mental state or private intent.

M EVALUATION SCOPE

The evidence concerns early target prediction within the evaluated Unity application, headset family, sensing stack, and non-locomotor task family. Zero-shot tests extend to two unseen task flows and multiple layouts within this platform; the same-event K-sweep characterizes sensitivity to candidate-list composition. Query gains are positive in aggregate across all evaluated horizons, with the strongest effect at 500 ms, while the crossed-summary and task analyses specify where those gains arise. The two-angle gaze correction and frozen selective policy provide deployment support, evaluated through drift tests and counterfactual replay. The APS interface is empirical for this repeated-measures setting, as specified in Appendix E.3. The evaluated scope is within-platform transfer and offline assistance.