CHRONOREPLAY: ONE ANSWER INTERFACE IS NOT ENOUGH FOR STREAMING SPATIAL INTELLIGENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Training a spatial VLM to answer which side an object lies on does not require it to preserve the metric displacement that determines that answer, while supervision for a distance estimate may leave categorical relations largely unconstrained. When post-training emphasizes one readout, improvements can therefore concentrate on that interface while transfer to others remains weak or even degrades; we call this observed specialization *interface lock-in*. ChronoReplay addresses it by replaying heterogeneous spatial supervision as chronologically ordered episodes, retaining source-native targets, and balancing supervision across families. Under the same observation access, a single ChronoReplay checkpoint improves OVO-S-Bench, OST-Bench, and UCS-Bench together, including a 13.41-point gain on OST-Bench, and achieves the highest three-benchmark mean among the supervision settings compared here. Same-data controls further separate the ingredients of this transfer: answer serialization redistributes gains across OST and UCS, family-balanced learning improves over ordinary joint SFT and natural turn weighting, and chronological replay exceeds a shifted-order control by about 2.41 OST points across two paired seeds. Adaptive frame selection then changes evidence access while keeping the trained model fixed, providing an additional inference-time gain at lower visual cost. **Streaming spatial performance depends on what distinctions supervision preserves, how heterogeneous histories are organized during learning, and what causal evidence inference makes available at test time.**

1 INTRODUCTION

A home robot follows you through your apartment and is later asked: *Is the mirror on my left or right? How far is the sofa from here? When did we pass the plant?* One walk, three questions, three answer forms: an option, a number in meters, a moment in time. Answering them from a video stream means associating objects across views, accumulating the observer's motion, and recovering relations in changing reference frames (Yang et al., 2025a; Plizzari et al., 2025; Yang et al., 2026b; Wang et al., 2026a; Thompson et al., 2025), using only what was observed before the question; streaming spatial benchmarks now pose all three over one causal history.

Recent progress comes from two directions. Architectural work retains the stream with visual memories, bounded caches, fast weights, maps, and scene graphs (Song et al., 2024; Shen et al., 2025; Liu et al., 2026; Gu et al., 2026); data work post-trains VLMs on spatial instruction data (Wu et al., 2025a; Xu et al., 2026; Yang et al., 2026a). Both are typically *single-interface*: a recipe supervises one answer format and reports the benchmark that shares it; even recipes that mix qualitative and metric questions (Chen et al., 2024a; Cheng et al., 2024) train on single images, so whether one model can serve categorical, metric, and temporal queries over the same causal history is untested. Building on this, we show that changing the supervision can achieve this without a new memory module or new annotation.

Post-training on individual supervision families exposes two challenges. On the one hand, ❶ *training can improve its dominant readout while transferring unevenly to other native readouts*. Starting from Qwen3-VL-8B, the categorical-only control reaches 58.03 on OVO-S but lowers OST from 50.52 to 47.48; trajectory-only training leaves OST nearly flat at 51.03. We use *interface lock-in* to name this empirical specialization across spatial readouts. The control motivates joint training, while

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.

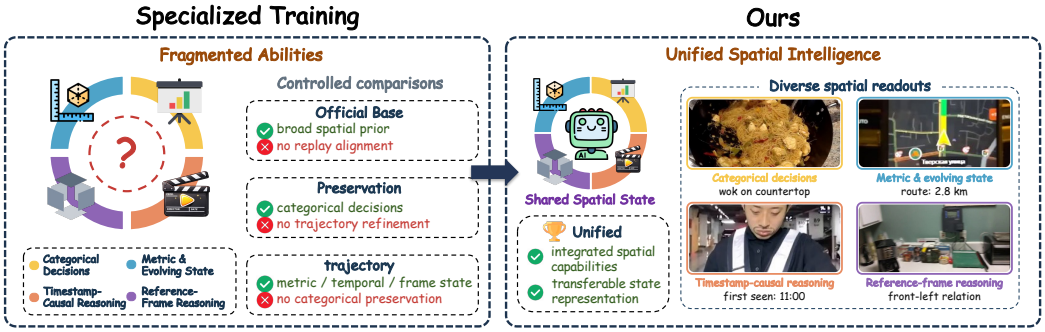


Figure 1: **Why reusable spatial state.** Categorical preservation and trajectory supervision constrain complementary spatial distinctions. Causal Replay combines them in shared parameters. The illustrated readouts span categorical, metric, temporal, and reference-frame queries.

differences in source and task coverage limit attribution to answer format alone (Appendix D.7). On the other hand, ❷ *a metric or temporal target is valid only under the history that produced it*. An observer that moves forward after facing north and one that does so after facing east share the same recent observations and the same local displacement, yet a later question about a previously seen object has different answers. Supervision that detaches a target from its valid prefix can therefore be ambiguous; it is crucial to keep each target attached to the chronological prefix under which it holds.

In this paper, we propose ChronoReplay, which consists of two components. **First**, to address ❶, *family-balanced learning* keeps every target in its *native format* (option semantics, units and precision, or timestamp) instead of collapsing it into one canonical answer type, and balances the two families in one objective to regulate their relative influence on the shared model. **Second**, to address ❷, *Causal Replay construction* replays both supervision families as chronologically ordered episodes in which every target is attached to the causal prefix under which it is valid, so an accumulated displacement or a retrieved timestamp is learned together with the origin and motion that give it meaning (*chronological alignment*). The main contributions are:

- We document an empirical *interface lock-in* pattern (❶): single-family training transfers unevenly across spatial readouts.
- To address ❶ and ❷, we propose ChronoReplay: family-balanced learning trains all interfaces in their native formats under one objective, and Causal Replay construction keeps every target attached to its chronological prefix. ChronoReplay gains +13.41 OST-Bench, +7.44 UCS-Bench, and +4.91 OVO-S-Bench points under identical observation access, with OVO-S numerically comparable to a published 235B reference.
- Matched interventions show a repeatable chronological-order benefit and a task-dependent answer-format tradeoff; adaptive frame selection at inference adds a separate gain at lower visual cost.

2 Related Work

Streaming spatial intelligence. Streaming and long-video models process observations causally and keep history in compressed visual memories, hierarchical summaries, recurrent memory bridges, active response mechanisms, or bounded key–value caches (Chen et al., 2024b; Ren et al., 2024; Qian et al., 2024; Zhang et al., 2025; He et al., 2024; Wang et al., 2025; Shu et al., 2025; Zhang et al., 2024; Huang et al., 2025; Qian et al., 2025; Chen et al., 2025; Xiong et al., 2025; Yang et al., 2025b; Zeng et al., 2025); StreamingBench and OVO-Bench formalize timestamped evaluation under causal access (Lin et al., 2026; Niu et al., 2025), and SimpleStream characterizes the trade-off between recent perception and historical retention (Shen et al., 2026). Spatial benchmarks further require history to stay coherent under object and observer motion and changes of reference frame, from 3D question answering, embodied spatial tests, and long-form egocentric video (Azuma et al., 2022; Ma et al., 2023; Du et al., 2024; Wang et al., 2024a; Mangalam et al., 2023) to multi-view, 4D, and mental-model probes (Li et al., 2025; Wang et al., 2026b); OVO-S-Bench, OST-Bench, and UCS-Bench extend them to causal streams with categorical, continuous, and timestamp-dependent

queries. In this work, we hold observation access fixed and treat it as an orthogonal factor (Section 5.4), and we use these three benchmarks because their answer formats differ; the recipe adds no memory mechanism, and our evaluation covers indoor scenes, so other environments remain untested.

Learning reusable spatial representations. Spatial MLLMs acquire geometry through geometry-aware visual features, multi-view or 3D-reconstruction inputs, and large-scale spatial instruction or reinforcement data (Cai et al., 2025; Liu et al., 2025; Zheng et al., 2025; Qi et al., 2026; Fan et al., 2026; Ouyang et al., 2025; Wu et al., 2025b), and persistent spatial state has been represented through 3D scene tokens and embodied generalist agents (Hong et al., 2023; Huang et al., 2024), adaptive fast weights, voxelized allocentric maps, textual scene descriptions, evolving scene graphs, and action-conditioned belief histories (Hua et al., 2026; Liao et al., 2026). Some mix qualitative and metric spatial questions on single images, and none that we know of asks whether categorical, metric, and temporal formats interfere when one checkpoint is trained on all of them over one causal stream. This is a multi-task question: which distinctions does a shared representation keep when objectives compete for the same parameters (Caruana, 1997; Sener & Koltun, 2018; Standley et al., 2020; Zhang & Yang, 2022), how should their losses be balanced, and what does fine-tuning on one distribution erase from another (Kirkpatrick et al., 2017; Luo et al., 2025; Zhai et al., 2024)? Instruction-tuning mixtures raise the same question at the level of data (Wei et al., 2022; Longpre et al., 2023; Liu et al., 2023), and the answer format itself changes what a model appears to know. In this work, we study it at the level of the supervision itself, through the chronology of training episodes and the format and weighting of their targets, and measure the outcome through native benchmarks.

3 Problem Formulation

At time t a model receives the causal observation history $h_t = (o_1, \ldots, o_t)$. A spatial query q arrives with a native answer interface $\kappa \in \mathcal{K}$ (a categorical option set, a metric value with units, or a timestamp), and the VLM predicts

$$z_t = f_\theta(h_t), \qquad \widehat{y}_t^{(\kappa)} = g_\theta(z_t, q, \kappa), \tag{1}$$

against the task-defined target $y_t^{(\kappa)} = A_\kappa(h_t, q) \in \mathcal{Y}_\kappa$. We call the representation z_t *reusable* over a family of query–interface pairs when every native target in that family can be read out from the same z_t. Reusability is a functional requirement, not a structural one: the state may stay distributed across the VLM, but histories that lead to different answers must remain distinguishable.

Two properties of spatial histories make this demanding. First, interfaces refine one another. A categorical target separates spatial regions, whereas a metric target for the same quantity resolves variation within a region, so a representation that keeps only the coarse partition cannot support the finer readout. Second, the state depends on the origin of the history. An observer that moves forward after facing north and one that moves forward after facing east share the same recent observations and the same local displacement, yet a later question about a previously seen object requires different answers. Histories with identical recent suffixes but different origins must therefore map to different states. Our goal is to learn one f_θ with both properties from heterogeneous native supervision; Section 4 builds training episodes that expose both.

4 ChronoReplay

4.1 Overview

ChronoReplay post-trains a VLM so that one shared representation supports native spatial readouts while preserving the history each readout depends on. Figure 2 summarizes two stages. *Causal Replay construction* forms the replay pool: every target is kept in its native format and every episode is ordered chronologically. *Family-balanced learning* normalizes the single-turn preservation and metric–temporal trajectory risks, then controls their relative contribution to the shared parameters. Algorithm 1 in Appendix C lists both stages.

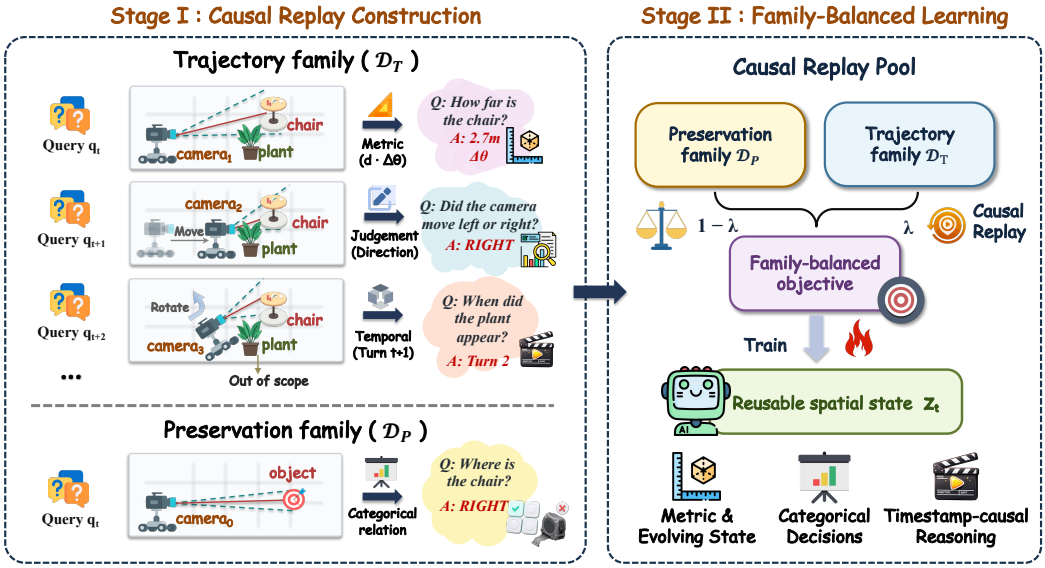


Figure 2: **Overview of ChronoReplay.** Stage I (Causal Replay construction) builds ordered trajectory supervision $\mathcal{D}_T$ and single-turn preservation examples $\mathcal{D}_P$. Stage II (family-balanced learning) combines their normalized native-format risks with weights $1-\lambda$ and λ. *Out of scope* marks observations beyond a query's causal frontier.

4.2 Causal Replay Construction

4.2.1 Native-Format Targets

Motivation. *How should heterogeneous answer spaces be combined?* Considering ❶, we keep each target in the value space of its source interface, avoiding an additional answer-space transformation when combining categorical options, metric quantities with units and precision, and temporal answers. This is a design choice for retaining source task semantics, rather than a claim that native encoding is uniformly superior. The format through which a model answers can change what it appears to know (Wang et al., 2024b) and the decoder can emit metric values directly (Vacareanu et al., 2024).

Native readouts. Let $F_\theta(C)$ denote the VLM representation of a causal prefix C (Section 3). For interface κ, let Y^κ denote its native target and G_κ the readout applied to $Z = F_\theta(C)$. Its population risk is

$$\mathcal{R}_\kappa(F_\theta, G_\kappa) = \mathbb{E}\left[-\log G_\kappa\left(Y^\kappa \mid Z, Q\right)\right]. \tag{2}$$

For a fixed representation and a sufficiently expressive readout, the optimal risk equals the conditional entropy $H(Y^\kappa \mid Z, Q)$, so a low risk requires Z to retain the information needed to recover the native target. Proposition 1 in Appendix A makes this precise. Corollary 1 shows that lossless re-encoding preserves this information requirement. Our native/canonical control changes output serialization without losing target content and finds crossed OST/UCS transfer effects.

Retained value spaces. Categorical targets keep their candidate structure; metric targets keep magnitude, sign, units, and annotation precision; temporal targets keep timestamps, event indices, and transition-relative quantities; reference-frame information stays coupled to the query. All targets are produced by the original autoregressive decoder.

4.2.2 Chronologically Aligned State Trajectories

What does a metric or temporal target mean without its history? With each target's format fixed, we now fix its history. Considering ❷, the same local displacement yields different world states from different origins (Section 3), so a target detached from its prefix is ambiguous. We use *temporal-origin alignment* to denote a data-construction rule: supervision from the interfaces present in a source episode is placed on that episode's video-time axis and replayed in its original order, with

each target attached to its legal causal prefix. Unlike flattening multi-round annotations into independent question–answer pairs, chronological replay preserves the source order and legal conditioning history used by the streaming evaluation protocol.

Ordered episodes. Let c_{ij} denote the legal causal prefix at turn j of episode i: all observations available at that turn together with the preceding interaction history. We represent an episode as

$$\mathcal{E}_i = (c_{ij}, q_{ij}, y_{ij}, \kappa_{ij})_{j=1}^{m_i}, \qquad c_{ij} \preceq c_{i,j+1}, \tag{3}$$

where $\preceq$ denotes causal prefix inclusion, q_{ij} is a spatial query, κ_{ij} identifies its native interface, and $y_{ij} = A_{\kappa_{ij}}(c_{ij}, q_{ij})$ is the target valid at that frontier. Single-turn preservation examples correspond to $m_i = 1$, whereas multi-round metric–temporal examples form longer episodes. Because the prefixes are nested, one episode yields m_i supervised states that share an origin and grow in history length, so the model is trained on accumulation itself, on how a displacement or an event order changes from turn to turn, rather than on isolated snapshots.

Implicit state trajectory. Applying the VLM to successive prefixes induces

$$z_{ij} = F_\theta(c_{ij}), \qquad \mathcal{Z}_i = (z_{i1}, \ldots, z_{im_i}). \tag{4}$$

The trajectory stays implicit within the VLM, and its first prefix defines the episode origin. No map, memory module, or auxiliary state loss is added. The origin-shift control in Section 5.3 changes the ordering and preceding context while retaining each turn block and its local supervision. It tests the contribution of this construction rule as a whole; it does not identify a particular latent origin representation. Proposition 2 in Appendix A characterizes the risk when the changed context makes the target ambiguous.

4.3 Family-Balanced Learning

Motivation. *What if the two families differ in size?* With targets and histories fixed, the remaining choice is how much each family may move the shared parameters. The preservation family $\mathcal{D}_P$ holds 33,180 single-turn examples (30,635 categorical and 2,545 numerical), whereas the trajectory family $\mathcal{D}_T$ holds 6,381 ordered episodes (45,006 supervised metric, counting, judgement, and temporal turns). Raw row counts could let one family dominate the shared parameters and reproduce ❶ inside joint training. Therefore, we normalize risk within each family and set their ratio explicitly rather than learning it (Kendall et al., 2018; Chen et al., 2018).

Family risks. Let ℓ_r^P be the native loss for preservation example r, with source-defined weight w_r (Appendix C). Let ℓ_{ij}^T be the native loss for verified trajectory transition (i, j), and define $\mathcal{T} = \{(i, j) : 1 \le i \le |\mathcal{D}_T|,\ 1 \le j \le m_i\}$. The two family risks are

$$\mathcal{L}_P = \frac{\sum_{r \in \mathcal{D}_P} w_r \ell_r^P}{\sum_{r \in \mathcal{D}_P} w_r}, \qquad \mathcal{L}_T = \frac{\sum_{(i,j) \in \mathcal{T}} \ell_{ij}^T}{|\mathcal{T}|}. \tag{5}$$

Balanced objective. ChronoReplay combines the risks as

$$\mathcal{L}_{\text{replay}} = (1 - \lambda)\, \mathcal{L}_P + \lambda\, \mathcal{L}_T, \qquad \lambda = \frac{\alpha}{1 + \alpha}, \tag{6}$$

where α is trajectory loss mass relative to preservation. We use $\alpha = 0.5$ in the main experiments: trajectory receives $1/3$ of total mass; preservation receives $2/3$, split equally between categorical and numerical targets. Appendix D specifies token normalization and serialized weighting; Proposition 3 states their invariances. Table 9 compares ordinary token SFT and natural turn weighting on the same replay pool.

Answer-length normalization assigns each supervised turn an average token loss, preventing long responses from dominating solely through token count. Within the trajectory family, turns receive equal weight, so an episode contributes in proportion to its supervised turns. Family weighting controls preservation and trajectory loss mass separately. This distinguishes response length, supervision frequency, and family balance within the training objective.

5 EXPERIMENTS

In this section, we evaluate ChronoReplay from multiple perspectives and organize the analysis around four research questions. **RQ1:** Does ChronoReplay transfer across categorical, continuous, and timestamp-causal benchmarks under fixed observation access (Section 5.2)? **RQ2:** How do single-family and joint training transfer across spatial readouts (Section 5.3)? **RQ3:** Which components carry the effect: native formats, chronological order, or family mass (Section 5.3)? **RQ4:** How much of the OVO-S system score is observation access rather than training (Section 5.4)?

5.1 EXPERIMENTAL SETUP

We evaluate on three benchmarks: OVO-S-Bench tests categorical reasoning on egocentric perception, object permanence, and allocentric mapping (Li et al., 2026); OST-Bench scores agent state, visible information, temporal localization, counting, and agent–object relations over multi-round exploration (Lin et al., 2025); and UCS-Bench has 8,114 timestamp-causal questions on trajectories, pose, proximity, object identity, and quantity (Wang et al., 2026c). All checkpoints start from the same official Qwen3-VL-8B-Instruct revision (Bai et al., 2025). ChronoReplay trains the language model and visual merger (visual encoder frozen) on the 33,180-sequence preservation family and the 6,381-episode trajectory family with $\alpha = 0.5$, and canonical scene identities keep every test scene out of training. The single-family diagnostic compares CATEGORICAL-ONLY and TRAJECTORY-ONLY variants with full ChronoReplay (Appendix D.7). Its categorical-only arm is distinct from the main preservation pool, which also contains numerical targets. SHIFTED cyclically rotates the turn blocks of every episode while keeping their content, controlling for chronological alignment; CANONICAL rewrites every target as a typed `value` field, controlling for native formats (the ChronoReplay default is NATIVE); and Uniform-81 matches the average frame count of adaptive access, an equal-cost control for observation access. Model-only comparisons hold causal observation access fixed (OVO-S Uniform-128, the complete scene-disjoint OST protocol, all UCS questions), while adaptive access (Appendix E) is an inference-time policy applied separately to the same checkpoint. Published models are reference points under their own protocols, and mechanism claims rely on matched Qwen3-VL-8B comparisons. Appendix D lists optimization settings, matched controls, and a four-GPU replication.

5.2 CROSS-INTERFACE TRANSFER (RQ1)

Table 1: **Complete task-family results on OVO-S-Bench** (%), twelve official task families. Our rows report accuracy within each family; Overall is their unweighted mean. Published reference scores are quoted from the benchmark paper. Uniform-128 is the matched training comparison; the final row applies adaptive access to the same trained model.

	Level 1			Level 2			Level 3			Level 4			Overall	Frames
Model	1.1	1.2	1.3	2.1	2.2	2.3	3.1	3.2	3.3	4.1	4.2	4.3		
Representative official prefix-access models														
InternVL-3.5-8B	47.86	54.60	35.31	40.95	40.31	56.22	54.00	26.98	60.61	34.65	55.00	28.30	44.57	128
Qwen3-VL-4B	32.91	53.19	43.65	47.26	34.25	63.04	57.33	38.10	68.18	25.74	70.00	28.30	46.83	128
Qwen3-VL-32B	48.83	57.07	44.51	50.71	34.58	70.43	60.67	26.98	68.18	32.67	68.33	22.64	48.80	128
InternVL-3.5-38B	58.64	55.56	49.74	50.30	48.00	64.91	61.33	22.22	53.03	38.61	68.33	18.87	49.13	128
InternVL-3.5-241B-A28B	61.30	59.36	45.98	48.75	49.86	68.54	53.33	33.33	68.18	42.57	65.83	13.21	50.85	128
GPT-5.4	52.08	63.70	48.13	46.25	57.92	68.61	61.33	36.51	54.55	32.67	62.50	26.42	50.89	128
Qwen3.5-27B	46.50	58.13	49.88	55.24	41.53	68.67	54.00	36.51	66.67	35.64	71.67	35.85	51.69	128
Qwen3.5-397B-A17B	46.71	53.32	48.68	55.65	41.22	69.30	57.33	38.10	78.79	34.65	75.00	26.42	52.10	128
Qwen3-VL-235B-A22B	50.36	56.09	50.91	49.76	45.61	70.17	63.33	46.03	74.24	45.54	70.83	20.75	53.64	128
Gemini-3.1-Pro	58.69	71.18	55.89	57.92	59.97	74.23	67.33	41.27	59.09	60.40	81.67	22.64	59.19	128
Our Qwen3-VL-8B evaluation under matched causal access														
Official Qwen3-VL-8B (Uniform-128)	57.49	60.31	44.40	54.55	47.02	70.07	59.33	26.98	62.12	19.85	61.11	24.53	48.98	128
ChronoReplay (Uniform-128)	62.87	64.43	48.51	59.09	51.79	74.15	65.33	33.33	66.67	24.43	67.78	28.30	**53.89**	128
ChronoReplay + adaptive access	69.46	71.13	61.19	60.10	58.33	76.87	68.67	34.92	68.18	26.72	70.00	30.19	57.98	80.50

Categorical streaming reasoning. Under matched Uniform-128 access, ChronoReplay raises OVO-S from 48.98 to 53.89. The score is numerically comparable to the published Qwen3-VL-235B-A22B reference. All twelve families improve; L4 rises from 35.16 to 40.17 but remains the

lowest-scoring level. The adaptive-access row is a system effect that Section 5.4 separates from training.

Table 2: **Fine-grained results on OST-Bench** (%), official fifteen subtypes. Published rows are reference points under their reported full-split protocols; shaded rows use our matched scene-disjoint evaluation. Δ is ChronoReplay minus base. JUD., TEMP., CNT., and EST. denote judgement, temporal localization, counting, and estimation.

	Avg	Agent State				Agent Visible Information					Agent–Object Spatial Relationship					
		Position		Orientation		Existence		Quantity	Diversity	Order	Direction			Distance		
Model		JUD.	EST.	JUD.	EST.	JUD.	TEMP.	CNT.	JUD.	JUD.	JUD.	TEMP.	EST.	JUD.	TEMP.	EST.
Representative published MLLMs																
InternVL-2.5-8B	44.6	51.7	26.1	49.4	40.4	86.3	51.3	56.4	60.7	38.4	37.2	33.8	22.8	43.0	42.9	27.9
Qwen2.5-VL-72B	45.6	57.0	27.6	52.2	37.1	86.1	64.5	61.5	75.7	34.5	41.4	21.1	8.2	44.5	39.3	18.7
Qwen2.5-VL-32B	46.9	51.0	31.1	53.5	39.4	85.3	64.8	59.2	73.4	41.8	39.5	24.9	25.7	43.6	39.1	20.3
Claude-3.5-Sonnet	47.8	65.1	36.2	50.6	30.3	85.8	67.0	57.9	57.1	60.0	39.2	18.3	21.9	43.8	54.9	19.0
GPT-4o	48.7	55.6	20.5	45.6	33.6	90.6	75.6	59.8	78.2	59.6	46.1	19.5	21.4	43.1	50.6	20.4
Gemini-2.0-Flash	49.5	59.7	27.3	56.7	36.5	89.6	70.8	59.4	78.7	55.6	42.0	17.8	21.5	45.5	48.9	27.1
InternVL-2.5-38B	50.8	56.7	31.8	54.6	38.4	91.7	74.7	61.1	79.8	62.1	42.1	20.6	27.7	42.7	42.5	28.1
InternVL-2.5-78B	51.1	60.8	34.4	49.9	40.7	90.7	74.4	65.9	77.9	61.2	43.4	22.4	17.8	46.7	44.4	22.9
GPT-4.1	53.4	64.2	30.7	60.8	33.2	90.8	78.0	60.6	82.1	70.8	51.5	23.4	28.6	44.9	53.6	23.9
Gemini-2.0-Flash (Thinking)	54.2	57.4	36.3	61.1	33.4	88.1	74.8	61.9	63.6	73.1	50.9	51.7	23.3	51.1	56.8	22.7
Our scene-disjoint Qwen3-VL-8B evaluation																
Official base	50.52	58.65	42.12	55.32	37.02	83.29	64.07	61.37	76.33	59.71	43.35	25.57	21.58	48.27	44.93	34.47
ChronoReplay	**63.93**	65.82	45.00	61.30	43.20	94.28	82.40	78.54	87.92	80.29	60.06	41.67	48.45	59.34	63.29	37.89
Δ	+13.41	+7.17	+2.88	+5.97	+6.18	+10.99	+18.33	+17.17	+11.59	+20.57	+16.71	+16.09	+26.86	+11.07	+18.36	+3.42

Continuous state estimation. On the scene-disjoint OST protocol, ChronoReplay improves OST from 50.52 to 63.93 (+13.41 points). The Δ row of Table 2 is positive in all fifteen subtypes. The largest changes are Direction Estimation (+26.86), Order Judgement (+20.57), and Distance Temporal Localization (+18.36). In short, gains are largest where an answer must be composed from accumulated multi-turn state rather than the current view, showing that ChronoReplay substantially improves continuous state estimation over multi-round exploration (RQ1).

Table 3: **Complete UCS-Bench decomposition** over all 8,114 timestamp-causal questions (%). Each group reports its aggregate (All) and two subcategories; category names follow Appendix B. Published rows are reference points under their reported evaluation protocols; Δ is ChronoReplay minus base.

			Trajectory			Position			Proximity			Category			Overall
Model	Size	Frames	All	My	Obj.	All	Ego	Rel.	All	Dist.	Reach	All	Cat.	Qty.	
Representative official references															
DeepSeek-V3	–	50	42.0	41.0	44.8	33.2	35.4	31.6	37.7	45.9	31.2	32.3	36.8	30.3	35.6
GPT-5	–	50	44.5	43.3	48.1	37.8	40.2	36.3	55.4	52.9	57.5	41.4	58.8	33.5	43.7
InternVideo2.5	8B	64	44.2	43.9	45.2	39.4	42.4	37.5	52.7	49.1	55.5	42.2	53.0	37.4	43.8
Qwen3-VL-8B	8B	64	44.6	43.1	48.7	40.1	43.6	37.9	50.9	49.2	52.1	42.9	52.8	38.4	44.0
Doubao-Seed-1.6V	–	50	49.6	49.4	50.1	42.8	46.4	40.4	56.0	54.9	56.8	44.6	57.0	39.0	47.3
Qwen2.5-VL-32B	32B	64	50.3	49.4	53.0	40.9	43.1	39.5	55.4	50.4	59.3	46.0	56.4	41.4	47.3
Qwen3-VL-32B	32B	64	50.3	49.0	54.1	46.5	50.1	44.3	55.2	53.9	56.3	46.7	58.8	41.2	49.0
InternVL3-38B	38B	64	53.7	52.4	57.3	45.6	46.9	44.8	55.3	54.0	56.3	45.8	60.4	39.3	49.2
DirectMe (Qwen3-VL)	8B	1fps	54.1	53.0	57.2	47.6	50.2	46.0	58.0	55.7	59.8	46.4	53.7	43.1	50.5
Human	–	–	91.7	90.9	94.1	87.6	84.0	89.8	96.8	95.0	98.2	90.0	94.3	88.1	91.0
Our Qwen3-VL-8B evaluation															
Official Qwen3-VL-8B	8B	128	48.49	47.49	51.49	41.72	40.32	42.61	49.49	46.12	52.14	43.87	53.50	39.60	45.33
ChronoReplay	8B	128	56.14	54.87	59.95	48.26	46.77	49.21	54.35	51.71	56.43	53.17	59.59	50.33	**52.77**
Δ			+7.65	+7.38	+8.46	+6.54	+6.45	+6.60	+4.86	+5.59	+4.29	+9.30	+6.09	+10.73	+7.44

Timestamp-causal reasoning. On UCS, ChronoReplay improves from 45.33 to 52.77 (+7.44 points), with positive transfer in all eight subcategories. The largest subcategory gains are on quantity (+10.73), object trajectory (+8.46), and observer trajectory (+7.38). These gains accompany categorical and metric improvements under unchanged observation access (RQ1).

Table 4: **Diagnostic analysis of ChronoReplay** (%). Panel (a): categorical-only, trajectory-only, and full ChronoReplay on Uniform-128 OVO-S and complete OST/UCS. ChronoReplay uses the canonical model from Tables 1–3; Mean averages the three benchmarks. Panel (b): each alternative is compared with its corresponding ChronoReplay reference under the scope shown. Order averages two seed pairs; the mass sweep uses adaptive OVO-S.

(a) Single-family supervision					(b) Mechanism controls			
Training	OVO-S	OST	UCS	Mean	Control: alternative	Alt.	ChronoReplay	Scope
Official base	48.98	50.52	45.33	48.28	Format: CANONICAL	58.03	**58.75**	OST/UCS
CATEGORICAL-ONLY	**58.03**	47.48	46.72	50.74	Order: SHIFTED	61.81	**64.22**	OST
TRAJECTORY-ONLY	52.03	51.03	49.22	50.76	Mass: $\alpha = 1.0$	60.92	**61.20**	OVO-S/OST
ChronoReplay	**53.89**	**63.93**	**52.77**	**56.86**	Topology: 4 GPUs	57.20	56.86	All three

5.3 DIAGNOSTIC ANALYSIS (RQ2, RQ3)

Key component analysis. Single-family training shows uneven cross-readout transfer; full ChronoReplay gives the highest three-benchmark mean. In the single-family diagnostic of Table 4(a), CATEGORICAL-ONLY reaches 58.03 on OVO-S while reducing OST to 47.48, whereas TRAJECTORY-ONLY reaches 52.03/51.03/49.22 on OVO-S/OST/UCS. The canonical ChronoReplay model reaches 53.89/63.93/52.77, yielding the highest three-benchmark mean of 56.86, compared with 50.74 and 50.76 for the two single-family variants. Source and task coverage differ between the single-family variants, so this comparison diagnoses cross-readout specialization rather than isolating answer interface as a single causal factor (Appendix D.7).

Comparison of joint-training objectives. With the joint data, chronological prefixes, native targets, and training budget held fixed, ordinary token SFT obtains 51.8/60.7/49.8 on OVO-S/OST/UCS, and natural turn weighting obtains 52.9/62.3/51.5 (Table 9). ChronoReplay improves all three endpoints, with a mean of 56.86 versus 54.10 and 55.57. This comparison supports the balanced objective beyond joint data exposure under the tested setting. Its OST advantage is 3.23 points over ordinary token SFT and 1.63 over natural turn weighting; the corresponding OVO-S/UCS gains are 2.09/2.97 and 0.99/1.27. The advantage over natural turn weighting repeats on every benchmark in each of three paired seeds (Table 10). Mean paired gains are 0.92 ± 0.27 on OVO-S, 1.43 ± 0.19 on OST, and 1.17 ± 0.25 on UCS (points, mean $\pm$ sample standard deviation).

Comparison of answer formats. With fixed response-only readers and matched four-GPU runs, NATIVE is higher on UCS by 2.92 points and in the two-benchmark mean (58.75 versus 58.03), while CANONICAL is higher on OST by 1.48 points. This crossed result shows that answer format redistributes transfer rather than improving every endpoint uniformly. We retain native targets to preserve source task semantics without an additional answer-space transformation; the control does not make native format a universally superior encoding. The canonical wrapper preserves the original answer exactly. By Corollary 1 in Appendix A, this is a lossless serialization intervention, with unchanged target information.

Comparison of training orders. Against SHIFTED, which keeps every turn block, target, and family weight, chronological replay is ahead in both seed pairs, by 2.42 and 2.41 OST points (mean 2.41). The shift changes both the episode origin and the context preceding each target; the result supports the temporal-origin alignment construction as a whole (RQ3). The 13.41-point OST improvement over the official base belongs to the complete recipe, while this paired ordering intervention measures a 2.41-point contribution from source chronology.

Sensitivity analysis. Over a fixed replay pool, $\alpha = 0.5$ gives the highest observed OVO-S–OST mean (61.20 versus 60.92 for $\alpha = 1.0$), and retraining on four GPUs instead of eight reproduces the three-benchmark mean within 0.34 points (RQ3; Appendix D).

Qualitative analysis. Figure 3 shows one episode whose two questions share a reference state: a temporal retrieval and a displacement relative to an earlier facing. ChronoReplay recovers both from the same origin, whereas SHIFTED retrieves a later round and flips the displacement, the failure the origin-shift control predicts (RQ3). Appendix F adds three more chronology cases, a mirror-history comparison using historical preservation and trajectory diagnostic models, and a fixed-observation

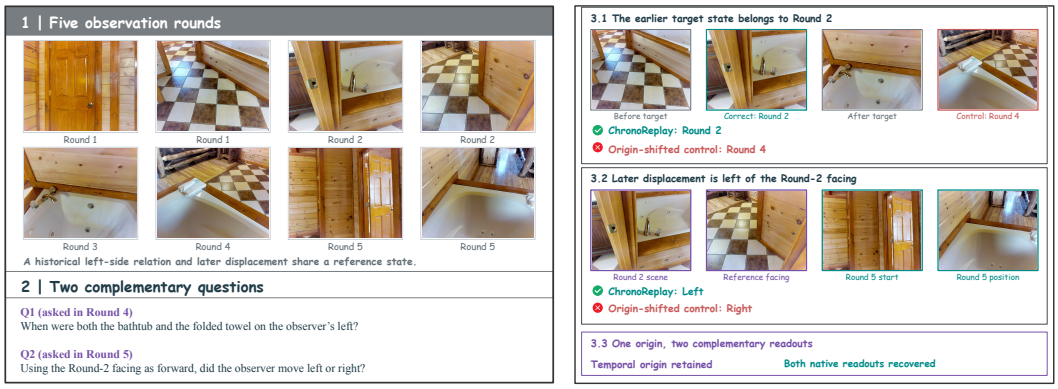


Figure 3: **One origin, two native readouts on an OST-Bench episode.** Left: the five observation rounds and two questions. Q1 (Round 4): when were both the bathtub and the folded towel on the observer's left? Q2 (Round 5): using the Round-2 facing as forward, did the observer move left or right? Right: ChronoReplay answers Round 2 and Left, both correct, whereas the origin-shifted control (SHIFTED) returns Round 4 and Right under identical observations. Teal marks annotated states, red the control's errors; the full-size episode is in Appendix F.

L4 layout failure (RQ4). The historical Preservation and Trajectory models used for qualitative analysis are specified separately in the appendix.

5.4 OBSERVATION ACCESS AS AN ORTHOGONAL FACTOR (RQ4)

Table 5: **Training–access decomposition on OVO-S** (%). Levels average their three family accuracies; Overall averages all twelve. Summary rows separate training and access effects, using unrounded scores.

Checkpoint	Access	L1	L2	L3	L4	Overall	Frames
Official base	Uniform-128	54.07	57.21	49.48	35.16	48.98	128
ChronoReplay	Uniform-128	58.60	61.68	55.11	40.17	53.89	128
ChronoReplay	**Adaptive**	67.26	65.10	57.26	42.30	**57.98**	80.50
Training effect (Uniform-128 − base)		+4.54	+4.46	+5.63	+5.01	+4.91	–
Access effect (adaptive − Uniform-128)		+8.66	+3.43	+2.15	+2.13	+4.09	–

Decomposition of the system score. Adaptive access adds +4.09 points beyond the +4.91-point training gain. Table 5 separates the two effects under the same scoring rule. Training improves all four levels, including L4 by 5.01 points. Adaptive access adds most on L1 (+8.66) and also improves L4 (+2.13), using 80.50 frames on average. L4 remains the hardest level. Appendix E.3 reports the base-access and Uniform-81 controls.

6 CONCLUSION

ChronoReplay replays heterogeneous spatial supervision in source chronology, retains native targets, and improves OVO-S, OST, and UCS together under matched observation access. We call the uneven cross-readout transfer observed in single-family training *interface lock-in*, an empirical pattern rather than an isolated internal mechanism. The single-family comparison retains the tradeoff: categorical-only training favors OVO-S, trajectory-only training gives little OST Overall gain, and full ChronoReplay has the highest three-benchmark mean. Objective controls show further gains beyond ordinary joint SFT. The full recipe gains 13.41 OST points over the base; the paired cyclic-shift control isolates a smaller, repeatable 2.41-point contribution from preserving source chronology as a whole. Native and canonical answer formats redistribute gains across UCS and OST rather than establishing a universally best encoding. OVO-S improves across all four levels, including +5.01 points on L4, which remains the hardest subset. These findings support ChronoReplay as a simple recipe for balanced cross-interface transfer without assigning its full gain to one component or claiming dominance on every spatial subcategory.

AI USE STATEMENT

Generative AI was used in two roles. AI assistants supported literature search, research brainstorming, manuscript revision, code editing, debugging, and data-pipeline implementation. In upstream data preparation, multimodal teachers were also used for bounded tasks such as proposing linguistic forms or assessing whether source evidence was visible and answerable. These teachers did not define spatial ground truth, and protected evaluation labels, predictions, and scores were unavailable to data construction. Every target used by ChronoReplay is bound to a source annotation, deterministic geometry, simulator state, or an audited public label; reported results are produced by frozen benchmark scorers. The authors designed the research, checked AI-assisted code, citations, claims, and prose against source papers and versioned experimental records, and take responsibility for the complete work.

ETHICS STATEMENT

This work uses existing research datasets and public benchmarks; it recruits no human participants and collects no new video or indoor scans. Source media may contain identifiable people, private-like interiors, or contextual information. Any artifact release therefore follows the corresponding source licenses and does not redistribute benchmark media. Canonical scene isolation prevents reported evaluation scenes and labels from entering training. Streaming spatial models may nevertheless inherit demographic, geographic, and environmental biases from their source data, and their ability to retain people, objects, and locations over time could enable surveillance or unsafe embodied decisions. We have not audited these high-stakes uses and do not claim suitability for surveillance, autonomous navigation, or other safety-critical deployment.

REPRODUCIBILITY STATEMENT

Appendices B–D specify the evaluation contracts, replay data, scene isolation, chronological construction, optimization, and controlled comparisons. Appendix E separately specifies the fixed-parameter observation-access intervention, and Appendix I records the executable configurations, generation parameters, scorer contracts, and source manifests. Every retained trainable comparison starts independently from the same pinned public Qwen3-VL-8B-Instruct revision; no intermediate checkpoint continuation or model fusion is used. Reported OVO-S, OST, and UCS scores bind model, data, protocol, prediction, and scorer identities, and all reported values are defined once as macros in `main.tex` and traced to versioned result records. We will release code, configurations, accepted training records or source manifests, prediction files, and scoring scripts subject to source licenses. Releasing the accepted records makes reproducing the reported training mixture independent of re-querying any external AI service.

REFERENCES

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. ScanQA: 3D question answering for spatial scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19129–19139, 2022.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitScenes: A diverse real-world dataset for 3D indoor scene understanding using mobile RGB-D data. In J. Vanschoren and S. Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. SpatialBot: Precise spatial understanding with vision language models. In *2025 IEEE International Conference on Robotics and Automation*, pp. 9490–9498, 2025. doi: 10.1109/ICRA55743.2025.11128671.

Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997. doi: 10.1023/A:1007379606734.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. In *International Conference on 3D Vision*, pp. 667–676. IEEE, 2017.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024a.

Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. VideoLLM-online: Online video large language model for streaming video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18407–18418, 2024b.

Joya Chen, Ziyun Zeng, Yiqi Lin, Wei Li, Zejun Ma, and Mike Zheng Shou. LiveCC: Learning video LLM with streaming speech transcription at scale. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29083–29095, 2025.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 794–803. PMLR, 2018.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. SpatialRGPT: Grounded spatial reasoning in vision-language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 135062–135093. Curran Associates, Inc., 2024.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2432–2443, 2017.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. EmbSpatial-Bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 346–355, Bangkok, Thailand, 2024. Association for Computational Linguistics.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Peihao Wang, Huaizhi Qu, Shijie Zhou, Dilin Wang, Zhicheng Yan, Hongyu Xu, Justin Theiss, Tianlong Chen, Jiachen Li, Zhengzhong Tu, Zhangyang Wang, and Rakesh Ranjan. VLM-3R: Vision-language models augmented with instruction-aligned 3D reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31054–31065, 2026.

Bo Gu, Zhikang Zhang, Zizhuang Wei, Zhenyuan Chen, Lingyun Li, and Zhuoyi Song. SpaceMind++: Toward allocentric cognitive maps for spatially grounded video MLLMs. *arXiv preprint arXiv:2605.09449*, 2026.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. MA-LMM: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13504–13514, 2024.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3D-LLM: Injecting the 3D world into large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 20482–20494. Curran Associates, Inc., 2023.

Jiacheng Hua, Yishu Yin, Yuhang Wu, Tai Wang, Yifei Huang, and Miao Liu. Unleashing spatial reasoning in multimodal large language models via textual representation guided reasoning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13616–13637, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.620.

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3D world. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20413–20451. PMLR, 2024.

Zhenpeng Huang, Xinhao Li, Jiaqi Li, Jing Wang, Xiangyu Zeng, Cheng Liang, Tao Wu, Xi Chen, Liang Li, and Limin Wang. Online video understanding: OVBench and VideoChat-Online. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3328–3338, 2025.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7482–7491, 2018.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Yifei Li, Pengyiang Liu, Yuhang Zang, Zhongyue Shi, Qi Fu, Hongye Hao, and Jiwen Lu. OVO-S-Bench: A hierarchical benchmark for streaming spatial intelligence in multimodal LLMs. *arXiv preprint arXiv:2606.03890*, 2026.

Yun Li, Yiming Zhang, Tao Lin, Xiangrui Liu, Wenxiao Cai, Zheng Liu, and Bo Zhao. STI-Bench: Are MLLMs ready for precise spatial-temporal world understanding? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5622–5632, 2025.

Chih-Ting Liao, Xi Xiao, Chunlei Meng, Zhangquan Chen, Yitong Qiao, Weilin Zhou, Tianyang Wang, Xu Zheng, and Xin Cao. SpaMEM: Benchmarking dynamic spatial reasoning via perception-memory integration in embodied environments. *arXiv preprint arXiv:2604.22409*, 2026.

Jingli Lin, Chenming Zhu, Runsen Xu, Xiaohan Mao, Xihui Liu, Tai Wang, and Jiangmiao Pang. OST-Bench: Evaluating the capabilities of MLLMs in online spatio-temporal scene understanding. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, volume 38, pp. 12769–12805, 2025.

Junming Lin, Zheng Fang, Chi Chen, Haoxuan Cheng, Zihao Wan, Fuwen Luo, Ziyue Wang, Peng Li, Yang Liu, and Maosong Sun. StreamingBench: Assessing the gap for MLLMs to achieve streaming video understanding. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 12147–12151, 2026. doi: 10.1109/ICASSP55912.2026.11463959.

Benlin Liu, Yuhao Dong, Yiqin Wang, Zixian Ma, Yansong Tang, Luming Tang, Yongming Rao, Wei-Chiu Ma, and Ranjay Krishna. Coarse correspondences boost spatial-temporal reasoning in multimodal language model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3783–3792, 2025.

Fangfu Liu, Diankun Wu, Jiawei Chi, Yimo Cai, Yi-Hsin Hung, Xumin Yu, Hao Li, Han Hu, Yongming Rao, and Yueqi Duan. Spatial-TTT: Streaming visual-based spatial intelligence with test-time training. In *European Conference on Computer Vision*, pp. 339–357. Springer, 2026.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. The flan collection: Designing data and methods for effective instruction tuning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22631–22648. PMLR, 2023.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 33:3776–3786, 2025.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. SQA3D: Situated question answering in 3D scenes. In *The Eleventh International Conference on Learning Representations*, 2023.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46212–46244. Curran Associates, Inc., 2023.

Junbo Niu, Yifei Li, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, and Jiaqi Wang. OVO-Bench: How far is your video-LLMs from real-world online video understanding? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18902–18913, 2025.

Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. SpaceR: Reinforcing MLLMs in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025.

Chiara Plizzari, Shubham Goel, Toby Perrett, Jacob Chalk, Angjoo Kanazawa, and Dima Damen. Spatial cognition from egocentric video: Out of sight, not out of mind. In *International Conference on 3D Vision*, pp. 1211–1221. IEEE, 2025.

Zhangyang Qi, Zhixiong Zhang, Ye Fang, Jiaqi Wang, and Hengshuang Zhao. GPT4Scene: Understand 3D scenes from videos with vision-language models. In *The Fourteenth International Conference on Learning Representations*, 2026.

Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 119336–119360. Curran Associates, Inc., 2024.

Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Dispider: Enabling video LLMs with active real-time interaction via disentangled perception, decision, and reaction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24045–24055, 2025.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. TimeChat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14313–14323, 2024.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. LongVU: Spatiotemporal adaptive compression for long video-language understanding. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 54582–54599. PMLR, 2025.

Yujiao Shen, Shulin Tian, Jingkang Yang, and Ziwei Liu. A simple baseline for streaming video understanding. *arXiv preprint arXiv:2604.02317*, 2026.

Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. Video-XL: Extra-long vision language model for hour-scale video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26160–26169, 2025.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. MovieChat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18221–18232, 2024.

Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9120–9132. PMLR, 2020.

Jacob Thompson, Emiliano Garcia-Lopez, and Yonatan Bisk. REM: Evaluating LLM embodied spatial reasoning through multi-frame trajectories. In *Second Conference on Language Modeling*, 2025.

Robert Vacareanu, Vlad-Andrei Negru, Vasile Suciu, and Mihai Surdeanu. From words to numbers: Your large language model is secretly a capable regressor when given in-context examples. In *First Conference on Language Modeling*, 2024.

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 75392–75421. Curran Associates, Inc., 2024a.

Pan Wang, Yang Liu, Guile Wu, Eduardo R. Corral-Soto, Chengjie Huang, Binbin Xu, Dongfeng Bai, Xu Yan, Yuan Ren, Xingxin Chen, et al. Spatial4D-Bench: A versatile 4D spatial intelligence benchmark. *arXiv preprint arXiv:2601.00092*, 2026a.

Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Jiajun Wu, Li Fei-Fei, and Manling Li. Spatial mental modeling from limited views. In *The Fourteenth International Conference on Learning Representations*, 2026b.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. "My answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 7407–7416, Bangkok, Thailand, 2024b. Association for Computational Linguistics.

Yun Wang, Junbin Xiao, Han Lyu, Yifan Wang, Jing Zuo, Zhanjie Zhang, Hong Huang, Dapeng Wu, and Angela Yao. Keep it in mind: User centric continual spatial intelligence reasoning in egocentric video streams. In *Forty-third International Conference on Machine Learning*, 2026c.

Yuxuan Wang, Yiqi Song, Cihang Xie, Yang Liu, and Zilong Zheng. VideoLLaMB: Long streaming video understanding with recurrent memory bridges. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24170–24181, 2025.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-MLLM: Boosting MLLM capabilities in visual-based spatial intelligence. In *Advances in Neural Information Processing Systems*, volume 38, pp. 15211–15239, 2025a.

Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. In *Advances in Neural Information Processing Systems*, volume 38, pp. 158982–159015, 2025b.

Haomiao Xiong, Zongxin Yang, Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Jiawen Zhu, and Huchuan Lu. Streaming video understanding and multi-round interaction with memory-enhanced knowledge. In *The Thirteenth International Conference on Learning Representations*, 2025.

Runsen Xu, Weiyao Wang, Hao Tang, Xingyu Chen, Xiaodong Wang, Fu-Jen Chu, Matt Feiszli, and Kevin J. Liang. Multi-SpatialMLLM: Multi-frame spatial understanding with multi-modal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31078–31088, 2026.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10632–10643, 2025a.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Rob Fergus, Yann LeCun, Li Fei-Fei, and Saining Xie. Cambrian-S: Towards spatial supersensing in video. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. MMSI-Bench: A benchmark for multi-image spatial intelligence. In *The Fourteenth International Conference on Learning Representations*, 2026b.

Yanlai Yang, Zhuokai Zhao, Satya Narayan Shukla, Aashu Singh, Shlok Kumar Mishra, Lizhu Zhang, and Mengye Ren. StreamMem: Query-agnostic KV cache memory for streaming video understanding. *arXiv preprint arXiv:2508.15717*, 2025b.

Xiangyu Zeng, Kefan Qiu, Qingyu Zhang, Xinhao Li, Jing Wang, Jiaxin Li, Ziang Yan, Kun Tian, Meng Tian, Xinhai Zhao, Yi Wang, and Limin Wang. StreamForest: Efficient online video understanding with persistent event memory. In *Advances in Neural Information Processing Systems*, volume 38, pp. 84409–84440, 2025.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the catastrophic forgetting in multimodal large language model fine-tuning. In Yuejie Chi, Gintare Karolina Dziugaite, Qing Qu, Atlas Wang Wang, and Zhihui Zhu (eds.), *Conference on Parsimony and Learning*, volume 234 of *Proceedings of Machine Learning Research*, pp. 202–227. PMLR, 2024.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, and Xiaojie Jin. Flash-VStream: Efficient real-time understanding for long video streams. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21059–21069, 2025.

Pan Zhang, Xiaoyi Dong, Yuhang Cao, Yuhang Zang, Rui Qian, Xilin Wei, Lin Chen, Yifei Li, Junbo Niu, Shuangrui Ding, Qipeng Guo, Haodong Duan, Xin Chen, Han Lv, Zheng Nie, Min Zhang, Bin Wang, Wenwei Zhang, Xinyue Zhang, Jiaye Ge, Wei Li, Jingwen Li, Zhongying Tu, Conghui He, Xingcheng Zhang, Kai Chen, Yu Qiao, Dahua Lin, and Jiaqi Wang. InternLM-XComposer2.5-OmniLive: A comprehensive multimodal system for long-term streaming video and audio interactions. *arXiv preprint arXiv:2412.09596*, 2024.

Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, 2022.

Duo Zheng, Shijia Huang, and Liwei Wang. Video-3D LLM: Learning position-aware video representation for 3D scene understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8995–9006, 2025.

A THEORETICAL ANALYSIS

This appendix formalizes the three arguments that motivate ChronoReplay (Sections 4.2.1, 4.2.2, and 4.3). Each result is stated at the population level with an unrestricted readout, so it describes what the objective asks of the representation; it does not certify that finite-data token optimization attains it. The controls in Section 5.3 examine the empirical relevance of these properties. Throughout, C is a causal prefix drawn from the replay pool, Q its query, κ the native interface, $Y^\kappa = A_\kappa(C, Q)$ the native target, $Z = F_\theta(C)$ the representation, and $H(\cdot \mid \cdot)$ and $I(\cdot;\cdot \mid \cdot)$ denote conditional entropy and conditional mutual information.

A.1 NATIVE READOUTS AS SUFFICIENCY CONSTRAINTS

Proposition 1 (Native risk and query-conditioned prediction)**.** *For any representation F_θ and any interface κ, the population risk of Eq. 2 satisfies*

$$\inf_{G_\kappa} \mathcal{R}_\kappa(F_\theta, G_\kappa) = H(Y^\kappa \mid Z, Q) = H(Y^\kappa \mid C, Q) + I(Y^\kappa; C \mid Z, Q), \tag{7}$$

where the infimum is over all conditional distributions of Y^κ given (Z, Q). For each query q in the support of Q, $H(Y^\kappa \mid Z, Q = q) = 0$ implies perfect prediction from (Z, q) on that conditional support.

Proof. For fixed (Z, Q), $\mathbb{E}[-\log G_\kappa(Y^\kappa \mid Z, Q)] = H(Y^\kappa \mid Z, Q) + \mathbb{E}\,\mathrm{KL}\big(P(Y^\kappa \mid Z, Q) \,\|\, G_\kappa(\cdot \mid Z, Q)\big)$, and the divergence vanishes exactly when G_κ equals the true conditional, which is admissible for an unrestricted readout; this gives the first equality. Since Z is a function of C, $H(Y^\kappa \mid C, Q) = H(Y^\kappa \mid C, Z, Q)$, and the chain rule gives $H(Y^\kappa \mid Z, Q) - H(Y^\kappa \mid C, Z, Q) = I(Y^\kappa; C \mid Z, Q) \geq 0$, which is the second equality. Zero conditional entropy means that the conditional target distribution is a point mass almost surely, yielding the query-conditioned prediction statement. □

We use this result to motivate supervision across heterogeneous query interfaces; cross-interface transfer is evaluated empirically in Section 5.

Corollary 1 (Lossless re-encoding preserves target information)**.** *Let $\tilde{Y} = g(Y^\kappa)$, where g is invertible on the target support. Then for every representation, $H(\tilde{Y} \mid Z, Q) = H(Y^\kappa \mid Z, Q)$ and $I(\tilde{Y}; C \mid Z, Q) = I(Y^\kappa; C \mid Z, Q)$. The two encodings impose the same target-sufficiency requirement at the population level.*

Proof. Each target is a deterministic function of the other. Conditional entropy cannot increase under either deterministic mapping, giving equality. Applying the conditional data-processing inequality in both directions likewise gives equality of the mutual information terms. □

Our canonical control wraps the original answer string in a JSON `value` field, so the target remains recoverable exactly. It tests transfer under a lossless change in output serialization. The crossed OST/UCS result in Table 4 therefore concerns finite-model training and readout behavior; it cannot be explained by a loss of target information in the wrapper. Equality at the population level does not imply identical token losses or optimization trajectories.

A.2 CYCLIC SHIFTS CHANGE THE CONDITIONING CONTEXT

Let O denote the identity of an episode's first turn block, C_j the legal chronological context for target j, and C'_j its preceding context after the full episode is cyclically shifted (Appendix D.5). The shift preserves each block's local content, but C'_j can omit earlier blocks and include blocks that originally occurred later. Knowing O therefore need not recover C_j from C'_j.

Proposition 2 (Risk under context ambiguity)**.** *Let $Y = A(C_j, Q)$ be a deterministic native target. With an unrestricted readout, the population log-loss floor under shifted conditioning is*

$$\inf_{G} \mathbb{E}[-\log G(Y \mid C'_j, Q)] = H(Y \mid C'_j, Q) = I(Y; O \mid C'_j, Q) + H(Y \mid C'_j, O, Q). \tag{8}$$

This floor exceeds $H(Y \mid C_j, Q) = 0$ if, on a set of shifted inputs (C'_j, Q) of positive probability, compatible chronological histories induce at least two distinct targets with positive conditional probability.

Proof. Expected log loss is conditional entropy plus the expected KL divergence from the true conditional distribution to G. The divergence vanishes at $G = P(Y \mid C'_j, Q)$, giving the first equality. The second is the chain rule for conditional mutual information. Under the stated ambiguity condition, the target distribution has positive entropy on a set of positive probability, so $H(Y \mid C'_j, Q) > 0$. Chronological conditioning has zero entropy because Y is a function of (C_j, Q). □

The residual term $H(Y \mid C'_j, O, Q)$ captures ambiguity that remains even when the origin is known, including ambiguity from missing history. Cyclic shifts need not satisfy the ambiguity condition: later observations introduced into a prefix may preserve or reveal the answer. The proposition therefore gives a conditional account of context corruption; it does not imply a strict penalty for every shift or determine downstream accuracy. The measured advantage of chronological replay (2.41 OST points averaged over two seeds in Table 12) supports preserving source order and legal causal prefixes together, without isolating a latent origin representation.

A.3 What Family Balance Does and Does Not Control

Proposition 3 (Invariances of the balanced objective)**.** *Let $\mathcal{L}_P$ and $\mathcal{L}_T$ be the family risks of Eq. 5 and $\mathcal{L}_{\text{replay}} = (1-\lambda)\mathcal{L}_P + \lambda\mathcal{L}_T$ with $\lambda = \alpha/(1+\alpha)$. Then: (i) replacing $\mathcal{D}_P$ or $\mathcal{D}_T$ by a family with the same within-family empirical distribution but a different number of sequences or supervised turns leaves $\nabla_\theta \mathcal{L}_{\text{replay}}$ unchanged, while α sets the explicit ratio of the two family weights; (ii) $\mathcal{L}_T$ is invariant to whether the supervised transitions of an episode are grouped into one sequence or presented as separate examples, provided each transition keeps its causal prefix; (iii) any common positive rescaling of all example weights, such as the one that sets the mean weight to one (Appendix D), leaves the minimizers of $\mathcal{L}_{\text{replay}}$ and the direction of its gradient unchanged.*

Proof. (i) Each family risk is a weighted mean, so it depends only on the empirical distribution of (ℓ, w) within the family and not on the number of terms; the same holds for its gradient. (ii) By Eq. 5, $\mathcal{L}_T$ is the plain average of ℓ^T_{ij} over all supervised transitions $(i,j) \in \mathcal{T}$, and each ℓ^T_{ij} is computed from the prefix c_{ij} alone (Eq. 11); grouping changes only how the same terms are batched. (iii) A common factor $s > 0$ on all weights cancels between the numerator and denominator of each family risk when applied before normalization, and multiplies the objective by s when applied after; in either case the arg-min and the gradient direction are unchanged. □

Proposition 3 is why the family-mass sweep of Table 11 is a sweep over a single scalar: with normalization in place, changing α changes the relative influence of the two families and nothing else. It also makes explicit what balance does not control: coverage of states within the trajectory family still grows with episode length, and no choice of α recovers information that a supervision family never supplied, which is the content of Proposition 1.

B Evaluation and Attribution Protocol

B.1 Attribution Framework

Performance depends jointly on the trained model and its observation access. Let $J(\theta, \pi)$ denote benchmark performance with model parameters θ and observation policy π. We separate the two effects as

$$\Delta_{\text{train}} = J(\theta^\star, \pi_0) - J(\theta_0, \pi_0), \qquad \Delta_{\text{access}} = J(\theta^\star, \pi_1) - J(\theta^\star, \pi_0), \tag{9}$$

where θ_0 is the official base model, $\theta^\star$ is the model trained with ChronoReplay, π_0 is the standard observation policy, and π_1 is adaptive access (Appendix E).

The first contrast changes the trained model under identical observation access. The second changes observation access while holding the model parameters fixed. The complete system evaluates $J(\theta^\star, \pi_1)$, while the two controlled contrasts separate the training and access effects.

We examine design choices through matched training interventions. The origin-shift intervention preserves the same multiset of turn blocks and supervision weights while changing their episode-level chronology and the accumulated context preceding each target. Initialization and optimization settings are held fixed. The target-interface experiment changes the answer representation while preserving the causal histories and training configuration. These comparisons probe sensitivity to the corresponding design choices of ChronoReplay.

B.2 Benchmark Protocols

OVO-S-Bench. Our fixed OVO-S evaluation uses only observations available by the query boundary. Uniform access to 128 frames defines π_0 and is used to measure Δ_{train}. Adaptive access defines π_1 for the complete system and selects one causal observation set before answer generation. All our OVO-S main results and component controls use the same scoring rule: family accuracy is $a_f = 100k_f/N_f$ and Overall is $\frac{1}{12}\sum_{f=1}^{12} a_f$. Each level averages its three families. We aggregate the unrounded fractions and round only the reported result. Table 6 gives the counts for the main results. The pooled accuracies are 50.91, 55.69, and 60.88 for Base, Uniform-128, and Adaptive, respectively; these are distinct from the reported family means of 48.98, 53.89, and 57.98.

Published external rows are quoted as references under their reported protocols.

Table 6: Correct answers underlying the OVO-S main results. Uniform-128 and Adaptive denote ChronoReplay; Base uses Uniform-128.

Family	N_f	Base	Uniform-128	Adaptive
1.1	167	96	105	116
1.2	194	117	125	138
1.3	268	119	130	164
2.1	198	108	117	119
2.2	168	79	87	98
2.3	147	103	109	113
3.1	150	89	98	103
3.2	63	17	21	22
3.3	66	41	44	45
4.1	131	26	32	35
4.2	90	55	61	63
4.3	53	13	15	16
Total	1,695	863	944	1,032

The access policy uses only the visible semantics of the query. Its routing rules and frame construction remain fixed throughout evaluation. The controlled comparison with uniform access to 81 frames measures the effect of temporal allocation at approximately equal visual cost.

OST-Bench. OST conversations are replayed in their original turn order. At each turn, the model receives the observations and interaction history available at that point in the episode. Evaluation uses a fixed split with disjoint training and test scenes, containing 8,351 turns in total and 7,814 turns scored by the official evaluator.

We report the complete official taxonomy: agent state, visible agent information, spatial relations between the agent and objects, judgement, temporal localization, counting, and estimation. The detailed table further decomposes these categories into the fifteen official question types.

UCS-Bench. We evaluate all 8,114 questions through the official interface. Each question is conditioned on observations available at its associated timestamp. The four official categories are Trajectory & Movement, Position & Orientation, Proximity & Reachability, and Category & Quantity, abbreviated as Trajectory, Position, Proximity, and Category in our tables. Each category has two subcategories: the twelve displayed scores comprise four category aggregates and eight subcategory scores (Wang et al., 2026c). All compared Qwen3-VL-8B checkpoints use the same observation and scoring protocol.

B.3 Scope of Comparison

Published systems provide external reference points under their reported observation settings. Attribution claims are based on matched Qwen3-VL-8B comparisons. Post-training is evaluated under fixed observation access, adaptive access is evaluated with fixed model parameters, and method components are evaluated through matched training interventions. This separation determines which conclusions can be drawn from each experimental block.

C Replay Supervision

C.1 Composition

ChronoReplay combines two supervision families. The single-turn preservation pool contains 30,635 categorical and 2,545 numerical examples. The trajectory pool contains ordered episodes with metric, counting, judgement, and temporal targets. The families differ in source and interaction structure as well as readout composition.

Table 7: Composition and relative loss mass of the Causal Replay pool at $\alpha = 0.5$. A supervised state is one causal prefix with a native target. Mass denotes the sum of training weights, normalized over the whole pool.

Family	Targets	Sequences	Supervised states	Loss mass
Preservation	Categorical	30,635	30,635	$1/3$
	Numerical	2,545	2,545	$1/3$
Trajectory	Mixed spatial	6,381	45,006	$1/3$

The complete data contain 39,561 serialized sequences. The trajectory family consists of 6,381 episodes and 45,006 supervised states. An episode may also contain observation turns that establish the spatial context for later targets. These turns enter the causal prefix and receive no direct loss.

Preservation sources and weights. The categorical pool comprises 9,000 source-annotated question–answer pairs, each presented with dense, contextual, and long-context observations (27,000 rows), plus 3,635 source-verified categorical examples under uniform causal observation access. The numerical pool comprises 945 ARKitScenes camera-rotation targets (Baruch et al., 2021) and 800 targets each for camera displacement and camera–object distance from corrected camera-geometry records. The trajectory pool uses the released OST training episodes.

Let $s = 13{,}187/12{,}635$. Each of the three views of a categorical pair has weight $s/3$, and each of the remaining 3,635 categorical examples has weight s. Every numerical example has weight $13{,}187/2{,}545$. A trajectory episode with m_i supervised turns has serialized weight $13{,}187\, m_i/45{,}006$ and averages its supervised turn losses. Thus the categorical, numerical, and trajectory weight sums are each 13,187, and the mean serialized-example weight is one. These are loss coefficients; they do not assert equal measured gradient norms.

The preservation and trajectory families retain their original questions and answer spaces. Categorical targets preserve candidate semantics, metric targets preserve magnitude and units, and temporal targets preserve event indices or timestamps. ChronoReplay therefore combines the two families through a shared model while maintaining the distinctions supplied by each source.

C.2 Episode Construction

Each trajectory is replayed in its original temporal order. At turn j, the model receives the observations available by that turn and all preceding interaction:

$$c_{ij} = (h_{ij}; q_{i1}, y_{i1}, \ldots, q_{i,j-1}, y_{i,j-1}) . \tag{10}$$

The current target y_{ij} follows this prefix as the next assistant response. Later observations and responses remain outside c_{ij}.

This construction preserves the trajectory origin and the state accumulated before each target. Object identity, observer pose, reference frame, and source units remain consistent throughout an episode.

The resulting data contain 47,096 turns, comprising 45,006 supervised states and 2,090 observation turns that provide context.

Keeping an episode intact also preserves dependencies across turns. A metric displacement is interpreted relative to the preceding observer state, while a temporal answer is interpreted relative to the events already observed. Training therefore associates each target with the causal context under which its spatial meaning is defined. Algorithm 1 summarizes the construction together with the family-balanced objective of Section 4.3.

Algorithm 1 ChronoReplay: Causal Replay construction and family-balanced learning.

Input: base VLM θ_0; preservation pool $\mathcal{D}_P$ with weights w_r; multi-round sources; trajectory mass α
Output: trained parameters θ

Stage I: Causal Replay construction
$\mathcal{D}_T \leftarrow \emptyset$
for each multi-round source episode i, turns kept in original order **do**
  for $j = 1, \ldots, m_i$ **do**
    $c_{ij} \leftarrow (h_{ij};\, q_{i1}, y_{i1}, \ldots, q_{i,j-1}, y_{i,j-1})$ ▷ causal prefix, Eq. 10
    keep y_{ij} in its native format; record its interface κ_{ij} ▷ Section 4.2.1
  end for
  $\mathcal{D}_T \leftarrow \mathcal{D}_T \cup \{(c_{ij}, q_{ij}, y_{ij}, \kappa_{ij})_{j=1}^{m_i}\}$ ▷ Eq. 3
end for
Stage II: family-balanced learning
$\mathcal{L}_P \leftarrow \sum_r w_r \ell_r^P / \sum_r w_r$; $\mathcal{L}_T \leftarrow |\mathcal{T}|^{-1} \sum_{(i,j)\in\mathcal{T}} \ell_{ij}^T$ ▷ Eq. 5
$\lambda \leftarrow \alpha/(1+\alpha)$
$\theta \leftarrow \arg\min_\theta (1-\lambda)\,\mathcal{L}_P + \lambda\,\mathcal{L}_T$, starting from θ_0 ▷ Eq. 6; one epoch, Table 8
return θ

C.3 Scene Separation

We separate training and evaluation by physical scene. A scene may appear under different frame rates, crops, filenames, or video encodings, so media files derived from the same scan are assigned a common scene identity. This identity is established before replay examples are constructed.

Our fixed OVO-S evaluation and the UCS evaluation share no physical scenes with the training data. For OST, we train on its released training data and evaluate on a fixed split with disjoint scenes. We withhold 243 overlapping ScanNet scenes (Dai et al., 2017), while retaining every official OST question type in the evaluation set.

Replay construction uses only annotations, observations, timestamps, reference frames, and answers from the training sources. Evaluation questions and answers play no role in example construction, family weighting, or target selection. The same scene separation applies to every model variant and matched intervention.

D Optimization and Controlled Comparisons

D.1 Training Settings

The primary model and its original controls are initialized independently from the same Qwen3-VL-8B-Instruct parameters. Training updates the language model and visual merger while keeping the visual encoder fixed. These original runs use one epoch of supervised learning with the configuration in Table 8. The single-family diagnostic in Appendix D.7 compares the canonical method with categorical-only and trajectory-only variants. The headline scores use one random seed. The natural-turn versus full-objective comparison covers three paired training seeds (Table 10). The comparison of chronological and shifted replay is repeated with two random seeds, with each pair sharing its initialization and optimization settings.

Table 8: Shared optimization settings for ChronoReplay and its matched controls.

Setting	Value
Base model	Qwen3-VL-8B-Instruct
Updated components	Language model and visual merger
Epochs / global batch size	1 / 32
Learning rate	2×10^{-6}
Schedule	Cosine decay with 3% warmup
Optimizer	AdamW, $\beta = (0.9, 0.999)$
Weight decay	0.01
Numerical precision	bfloat16
Maximum sequence length	32,768 tokens
Maximum processed pixels	200,704 per visual item
Relative trajectory mass	$\alpha = 0.5$

The same tokenizer, image processor, and conversation template are used across runs. Loss is applied to supervised assistant responses, with observations and preceding interaction providing causal context. Appendix I specifies the model revision and execution details.

D.2 Loss Normalization

We normalize supervision at the answer, transition, and family levels. For an answer sequence $y = (y_1, \ldots, y_N)$ following causal context c and query q, the native loss is

$$\ell_\theta(c, q, y, \kappa) = -\frac{1}{N} \sum_{n=1}^{N} \log p_\theta(y_n \mid c, q, \kappa, y_{<n}). \tag{11}$$

Averaging over answer tokens gives each target a contribution normalized by its length.

For episode i with m_i supervised turns, we average the turn losses and assign the episode a weight proportional to m_i, which is the same as giving every supervised transition one common weight within the trajectory family while retaining its complete causal context. Longer episodes therefore contribute in proportion to their number of supervised turns.

Preservation examples retain the weights supplied by their source. The total trajectory weight is then scaled to α times the total preservation weight, and a common rescaling sets the mean example weight to one. This implements the family risks defined in the main text.

For a fixed set of causal contexts and targets, this weighted empirical objective is invariant to grouping turns into episodes or separate examples. The grouping used during optimization determines how those contexts are presented.

Joint-training objective comparison. Table 9 compares three objectives on the same joint replay data, with chronological prefixes, native targets, training budget, and evaluation access held fixed. Ordinary token SFT uses the standard loss over supervised assistant tokens. Natural turn weighting retains answer-length normalization and within-family weighting, with family mass determined by the resulting turn weights. ChronoReplay additionally sets the relative family mass explicitly through $\alpha = 0.5$.

Table 9: Joint-training objective comparison (%). OVO-S uses Uniform-128; OST and UCS use the same fixed protocols as the main results. Mean is the unweighted average of the three benchmark scores.

Training objective	OVO-S	OST	UCS	Mean
Base	48.98	50.52	45.33	48.28
Ordinary token SFT	51.8	60.7	49.8	54.10
Natural turn weighting	52.9	62.3	51.5	55.57
ChronoReplay	**53.89**	**63.93**	**52.77**	**56.86**

Both ordinary token SFT and natural turn weighting improve all three benchmarks over the base. Family-balanced learning provides a further gain on each benchmark. These comparisons support the combined weighting design without separately isolating every level of normalization.

Paired-seed replication. Table 10 repeats the natural-turn versus full-objective comparison across three paired training seeds. The first pair corresponds to the original scores in Table 9; the other two extend the comparison to additional seeds. The full objective improves all three benchmarks in every pair. The table reports the sample standard deviation of the paired differences to describe variation across these runs.

Table 10: Three-seed replication of the objective comparison (%). Natural: natural turn weighting; Full: ChronoReplay; Δ: Full minus Natural in percentage points. Pair indices enumerate runs. Differences and summaries use the displayed scores; summary values are rounded after aggregation.

	OVO-S			OST			UCS		
Pair	Natural	Full	Δ	Natural	Full	Δ	Natural	Full	Δ
1	52.90	53.89	+0.99	62.30	63.93	+1.63	51.50	52.77	+1.27
2	52.71	53.33	+0.62	62.58	63.83	+1.25	51.68	52.56	+0.88
3	53.02	54.17	+1.15	62.14	63.56	+1.42	51.39	52.74	+1.35
Mean	52.88	53.80	**+0.92**	62.34	63.77	**+1.43**	51.52	52.69	**+1.17**
SD of Δ	–	–	0.27	–	–	0.19	–	–	0.25

D.3 Relative Weight of Trajectory Supervision

The family-mass sweep is a separate matched set of runs. We vary α while holding training examples, chronological order, initialization, and optimization settings fixed within this set. The $\alpha = 0.5$ row is an independent replication under the sweep protocol, distinct from the headline checkpoint. Table 11 reports OVO-S under the same adaptive observation policy and OST under the complete evaluation protocol. For comparison under adaptive access, the headline model scores 57.98 on OVO-S and 63.93 on OST. The final column is the arithmetic mean of the two benchmark scores.

Table 11: Sensitivity to the relative weight of trajectory supervision in a separate matched run set (%). OVO-S uses the fixed adaptive access policy.

α	OVO-S	OST	Mean
0.25	57.56	62.73	60.15
0.50	**58.27**	**64.13**	**61.20**
1.00	57.80	64.04	60.92

Among the tested settings, $\alpha = 0.50$ gives the highest score on both benchmarks. This comparison supports an intermediate allocation of supervision between single-turn preservation and trajectory learning.

D.4 Target Interface

This comparison examines how answer representation affects transfer. Two models share the same 39,561 training examples, initialization, random seed, global batch size, training schedule, and family weights. The native variant retains the original source responses. The canonical variant serializes each complete original answer string as {`"value": "original answer"`}, retaining units, precision, and wording. JSON decoding recovers the original target exactly. Observations, questions, supervised-turn positions, row order, and training weights are unchanged; the intervention changes the assistant's output serialization.

Evaluation converts each response into the input expected by the official scorer using a fixed reader for its trained format. Each reader operates on the generated response alone. All 8,351 OST responses and all 8,114 UCS responses yield an extracted value; the official OST scorer evaluates its designated 7,814 turns. Detailed extraction rules are provided in Appendix I. The comparison cov-

ers OST and UCS, whose generated answers can be evaluated through these readers. OVO-S uses greedy generation with deterministic option-label extraction.

Canonical supervision obtains 64.72 on OST and 51.34 on UCS. Native supervision obtains 63.24 and 54.26, respectively. Canonical targets improve OST by 1.48 points, while native targets improve UCS by 2.92 points. Native supervision consequently yields the higher mean across the two benchmarks. This crossed response indicates that the training interface affects the distribution of transfer across tasks.

D.5 Chronological Alignment

We examine the contribution of chronology by cyclically shifting the turn blocks within each episode. Let b_i denote the total number of turn blocks, including turns that provide observation context. For $b_i > 1$, a deterministic offset $d_i \in \{1, \ldots, b_i - 1\}$ defines

$$\sigma_i(j) = 1 + \big((j + d_i - 1) \bmod b_i\big). \tag{12}$$

The offset is determined by the episode identifier. Episodes containing one block retain their original order.

Each block carries its observations, query, response, and supervision weight into the shifted sequence. The two variants therefore share the same local content and total family weight. They also share initialization and optimization settings. The shift changes the starting point of the episode and the accumulated context preceding each target.

We repeat the comparison with two random seeds. Within each pair, chronological and shifted replay share the same initialization, training data, and optimization settings. Both models are evaluated on identical conversations in their original temporal order. Table 12 reports their OST performance.

Table 12: Chronological and shifted replay on OST (%). Each row compares models trained with the same seed and evaluated on the same scenes.

Training order	Run 1	Run 2
Cyclic shift	61.51	62.11
Chronological	**63.93**	**64.52**

Chronological replay is ahead by 2.42 and 2.41 points in the two comparisons, a mean advantage of 2.41 points. Because the shift also changes the context preceding each target, the comparison tests chronological alignment as a whole rather than the episode origin alone.

D.6 Training-Topology Replication

Table 13 tests whether the cross-interface result depends on the synchronized training topology. Both rows start from the same pinned official checkpoint and use the exact same 39,561 training rows, seed, optimizer, global batch of 32, and one-epoch schedule. The replication changes world size from eight to four and compensates by changing gradient accumulation from two to four.

Table 13: Fresh training-topology replication under complete model-only evaluation (%), together with the frozen OVO-S system protocol. Observation access and scorers are fixed within each reported column.

Run	GPUs	OVO-U128	OVO-Adaptive	OST	UCS	Model-only mean
Headline checkpoint	8	53.89	57.98	63.93	52.77	56.86
Fresh replication	4	54.10	58.00	63.24	54.26	57.20

The rerun preserves positive transfer over the official base on all three endpoints and improves the model-only mean by 0.34 points. Relative to the headline checkpoint, OVO-S Uniform-128 increases by 0.21 points, UCS by 1.49, and OST decreases by 0.69. This supports robustness to the distributed training topology, but does not constitute an independent-seed variance estimate. The fixed adaptive access policy raises the replica from 54.10 to 58.00, an access gain of 3.90 points and a close replication of the headline 57.98 system score.

D.7 Single-Family Supervision Control

Table 14 gives the single-family diagnostic summarized in Table 4(a). The ChronoReplay row is the canonical model reported in Tables 1–3. This categorical-only arm differs from the full preservation pool, which also includes numerical examples (Table 7).

Table 14: Single-family diagnostic on OVO-S, OST, and UCS (%). The ChronoReplay row uses the canonical model from Tables 1–3. Mean averages the three benchmarks under the evaluation protocols described above.

Training	OVO-S	OST	UCS	Mean
Official base	48.98	50.52	45.33	48.28
Categorical-only	**58.03**	47.48	46.72	50.74
Trajectory-only	52.03	51.03	49.22	50.76
ChronoReplay	**53.89**	**63.93**	**52.77**	**56.86**

Categorical-only training leads OVO-S but falls below the base on OST; trajectory-only gives little OST Overall gain. Both single-family arms improve UCS over the base. Canonical ChronoReplay achieves the highest three-benchmark mean, 56.86, compared with 50.74 for categorical-only and 50.76 for trajectory-only training. This comparison measures the observed specialization pattern across the single-family variants and the full recipe. The training families still differ in supervised tasks, so the table does not assign the full joint gain to one isolated mechanism.

E Adaptive Access: Temporal Observation Selection

Spatial queries differ in the history needed to answer them. Current relations can often be inferred from recent observations, whereas motion and event order depend on changes over time. Layout and historical references call for broader coverage. We use this distinction to adapt observation selection to the temporal scope of each question. The policy is fixed before evaluation and applied to the same trained model on OVO-S; OST and UCS retain their original observation protocols.

E.1 Matching Observations to Query Scope

Table 15 summarizes three complementary forms of temporal coverage. Local sampling concentrates on the current state. Multiscale sampling emphasizes recent changes while retaining earlier observations. Global sampling covers the complete history available at the query time.

Table 15: **Temporal coverage for different spatial queries.** Frame budgets are upper limits; short histories and duplicate removal can yield fewer frames.

	Local	**Multiscale**	**Global**
Query scope	Current state	Motion and event order	Layout and historical references
Coverage	Recent observations	Last 30 s; 30–300 s; earlier history	Full observed prefix
Sampling	Dense tail at 4 fps	60% / 30% / 10% of the frame budget	Uniform
Frame budget	8	128	128

Selection follows the visible question text. Named historical endpoints and persistent relative pose take priority and receive global coverage. Expressions of motion, change, or chronological order receive multiscale coverage. Explicit questions about the current state receive local coverage. Static layout and ambiguous temporal scope default to global coverage. A named historical reference therefore takes precedence over an incidental mention of the current state.

The selector uses question text alone. Answer choices, reference answers, benchmark categories, model predictions, and scores are excluded. Each question is answered once using the selected observations.

E.2 Sampling the Causal History

Let t_q be the query time and f the video frame rate. We retain the evaluated boundary convention $e = \max(1, \lfloor f t_q \rfloor)$, with frame indices starting at zero. The benchmark timestamps place this boundary within the video.

Local and global coverage. Global sampling takes the integer floors of 128 equally spaced positions on the inclusive interval $[0, e]$, then sorts and removes repeated indices. Local sampling uses spacing $d = \max(1, \mathrm{round}(f/4))$ and the index sequence

$$I_{\mathrm{local}} = \{\max(0, e - 7d), \max(0, e - 7d) + d, \ldots, e\}, \tag{13}$$

retaining the final eight indices if rounding produces an extra sample.

Multiscale coverage. With budget $B = 128$, the most recent 30 seconds receive $\mathrm{round}(0.6B)$ samples. The preceding interval extending to 300 seconds before the query receives $\mathrm{round}(0.3B)$ samples, and the remaining budget covers the earlier history.

Define $s_{30} = \max(0, e - 30f)$ and $s_{300} = \max(0, e - 300f)$. Let $\mathrm{Lin}(a, b, n)$ return the integer floors of n equally spaced positions in $[a, b)$. The requested indices are

$$\begin{aligned} I_{\mathrm{multi}} = {} & \mathrm{Lin}(s_{30}, e, \mathrm{round}(0.6B)) \\ & \cup \mathrm{Lin}(s_{300}, s_{30}, \mathrm{round}(0.3B)) \\ & \cup \mathrm{Lin}(0, s_{300}, n_{\mathrm{early}}), \end{aligned} \tag{14}$$

where n_{early} is the remaining budget. Empty intervals are skipped and repeated indices are removed. Global and local sampling include the query boundary; multiscale intervals exclude their right endpoint. These conventions preserve the sampling procedure used for the reported results.

E.3 Contribution of Observation Selection

Adapting temporal coverage benefits both the official model and ChronoReplay. The official model improves from 48.98 to 54.08, while ChronoReplay improves from 53.89 to 57.98 under our fixed OVO-S evaluation protocol. Each comparison holds the model parameters fixed and varies the observations presented at inference.

The comparison at a similar frame budget further examines temporal placement. Uniform sampling of 81 positions from the ordered set of 128 uniform indices obtains 52.40, whereas adaptive selection obtains 57.98 using 80.50 frames on average. The advantage of 5.58 points supports matching temporal coverage to query scope at approximately the same visual cost. Table 5 separates the training and access effects using the same 12-family mean.

F Qualitative Spatial Analysis

The first four cases examine three dependencies of native spatial answers: historical state retrieval, reference-frame transformation, and accumulated displacement; their scenes come from ScanNet, Matterport3D, and ARKitScenes (Dai et al., 2017; Chang et al., 2017; Baruch et al., 2021). Chronological and origin-shifted checkpoints receive identical ordered observations and questions at evaluation, under the comparison defined in Appendix D.5. Figures are selected examples of output differences; they do not estimate the prevalence of errors or identify an observation-selection failure. Figures 5–7 use Run 2 of Table 12.

Two additional cases examine native supervision complementarity and a global-layout failure. The first uses the historical diagnostic models specified in Appendix F.4; the second compares Base and the headline ChronoReplay model under fixed observations.

F.1 Historical State Retrieval

Figure 4 illustrates how an earlier state supports both a temporal answer and a later reference-relative displacement. The retrieval question requires the round in which the two objects occupied the specified side of the observer. The outputs distinguish recovery of that earlier state from selection of a later observation.

Figure 5 examines the related distinction between first appearance and category order. The queries require identifying which object was observed and when it first appeared. The apparent early match is a different fixture, making object identity part of the temporal retrieval problem.

F.2 Reference-Frame Transformation

Figure 6 combines a historical co-visibility query with a rotation query anchored to an earlier facing direction. The second answer depends on interpreting later observations in the specified reference frame. Its sign error is distinct from returning the wrong round for the first question, although both queries depend on earlier observations.

F.3 Accumulated Displacement and Historical Proximity

Figure 7 asks about displacement relative to a reference facing and about the closest previously observed state. The first requires tracking a location change, rather than reporting turning direction; the second requires comparing historical distances using the source annotation. The paired outputs illustrate two ways an answer can disagree with its historical reference under the same evaluation observations.

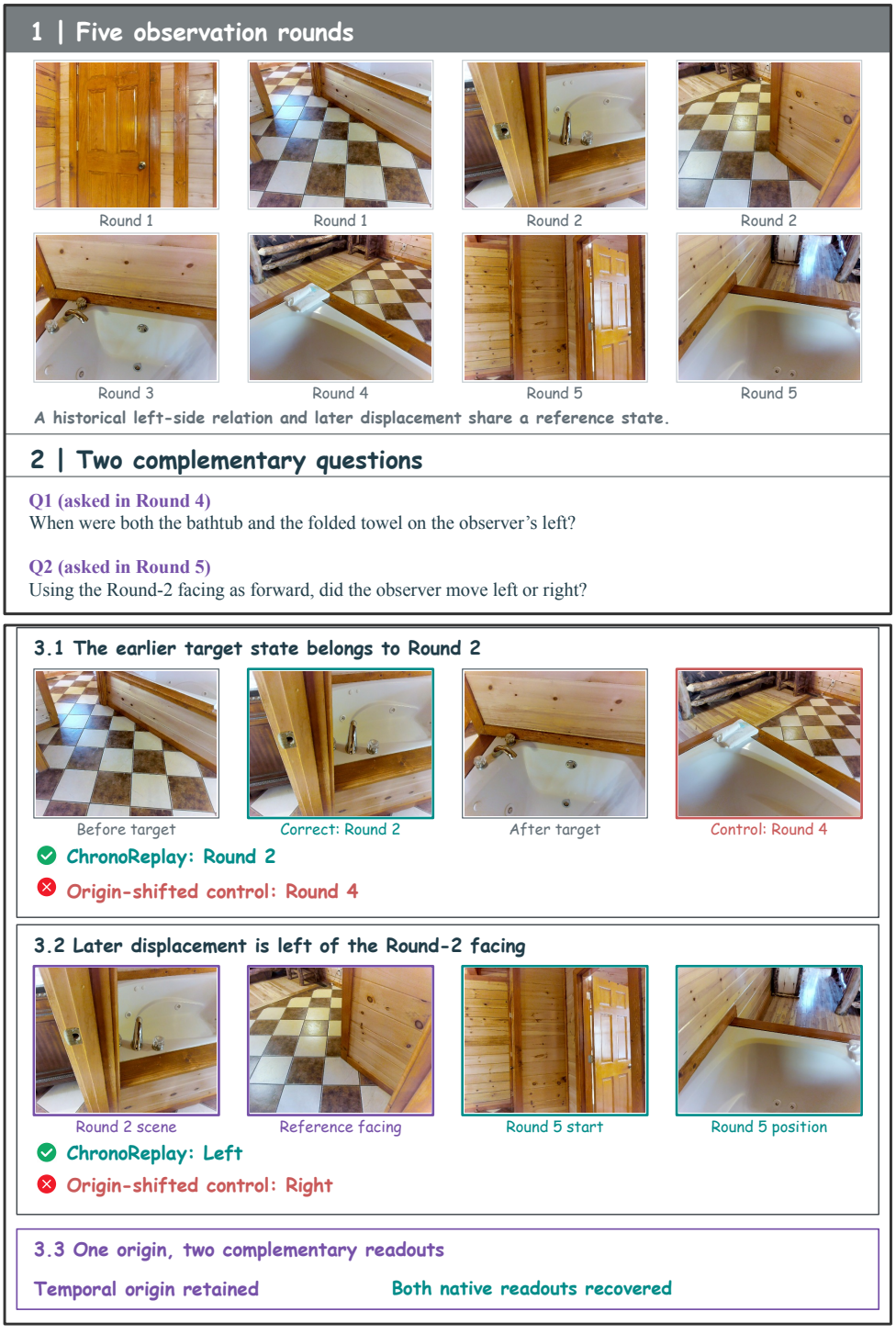


Figure 4: **Historical context in spatial reasoning.** The first question retrieves an earlier state in which the bathtub and folded towel lie on the observer's left. The second evaluates a later location relative to the facing direction in Round 2. Teal indicates annotated states, and red shows the control's incorrect retrieval or displacement. Both questions depend on a reference established earlier in the episode.

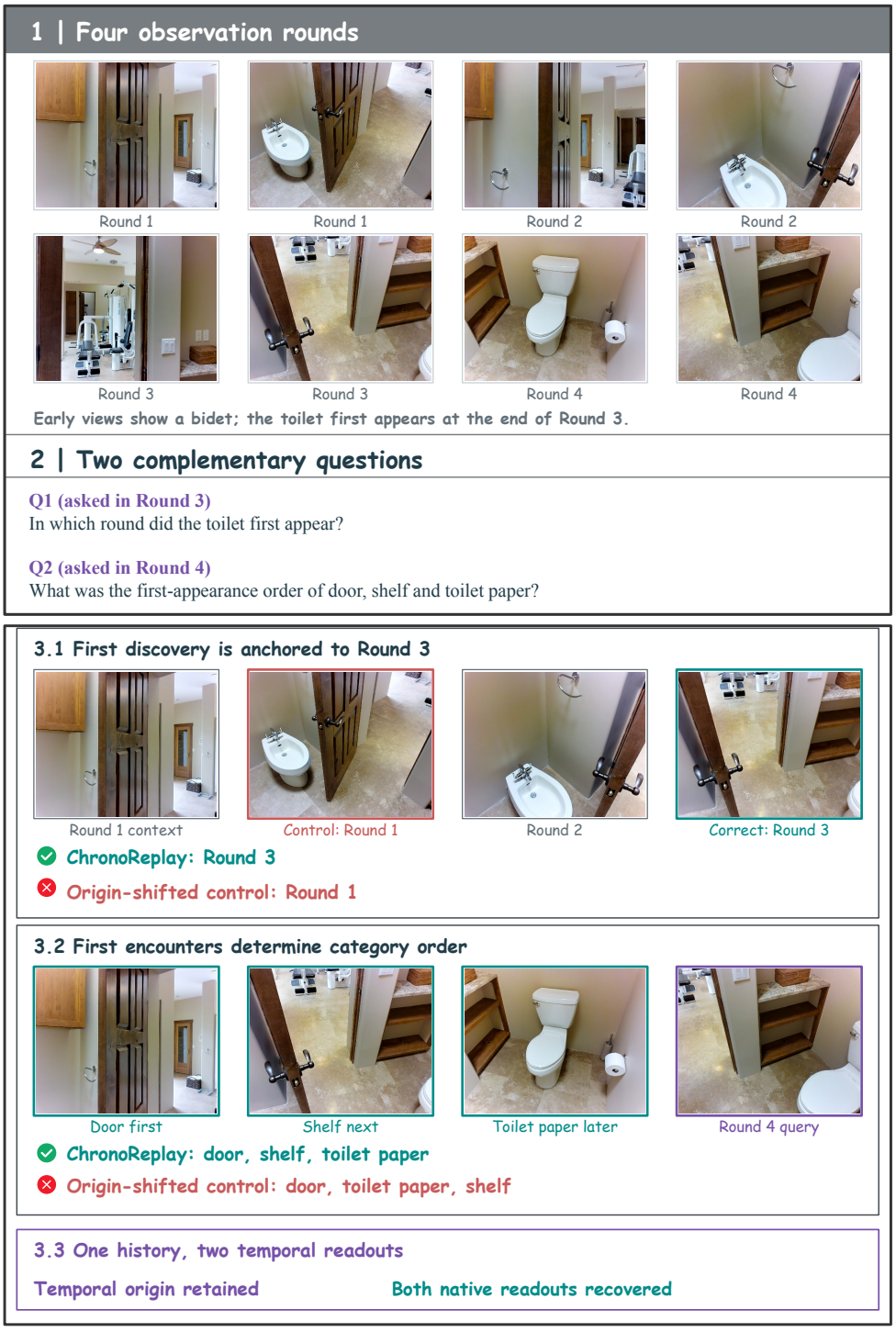


Figure 5: **First appearance and category order.** In scene `1mp3d_0067_region16`, ChronoReplay returns Round 3 for the first appearance of the toilet and the order door, shelf, toilet paper. The matched origin-shifted control returns Round 1 and reverses the latter two categories. The early white fixture is a bidet; the toilet is partially visible at the lower-right edge of the final Round 3 frame.

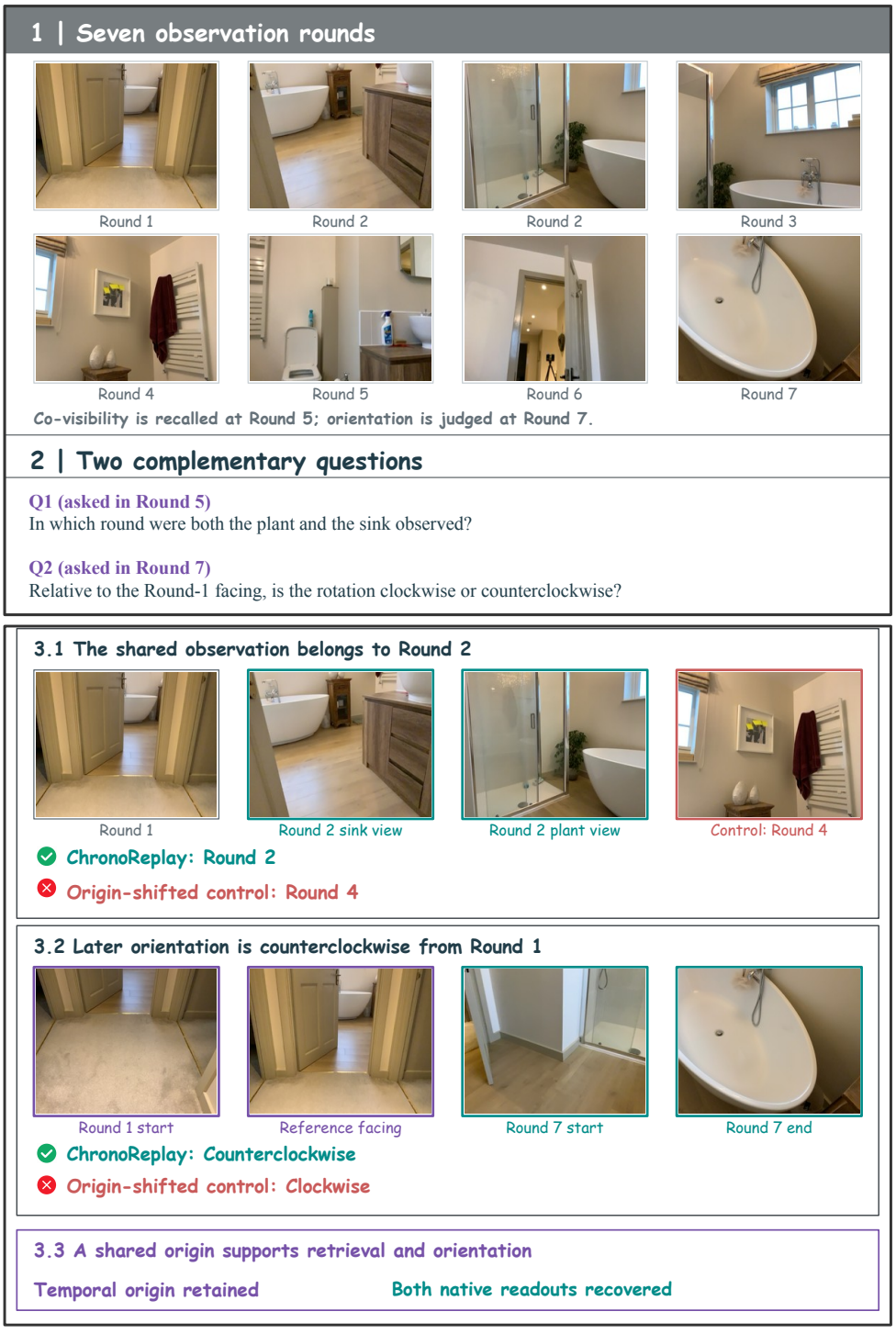


Figure 6: **Historical retrieval and rotation.** In scene `Validation_42898818`, the Round 5 question asks which round contained both the plant and sink; ChronoReplay returns Round 2, whereas the control returns Round 4. A Round 7 question defines the Round 1 facing as the orientation reference and specifies a rotation below 180°. ChronoReplay correctly returns counterclockwise, while the control returns clockwise.

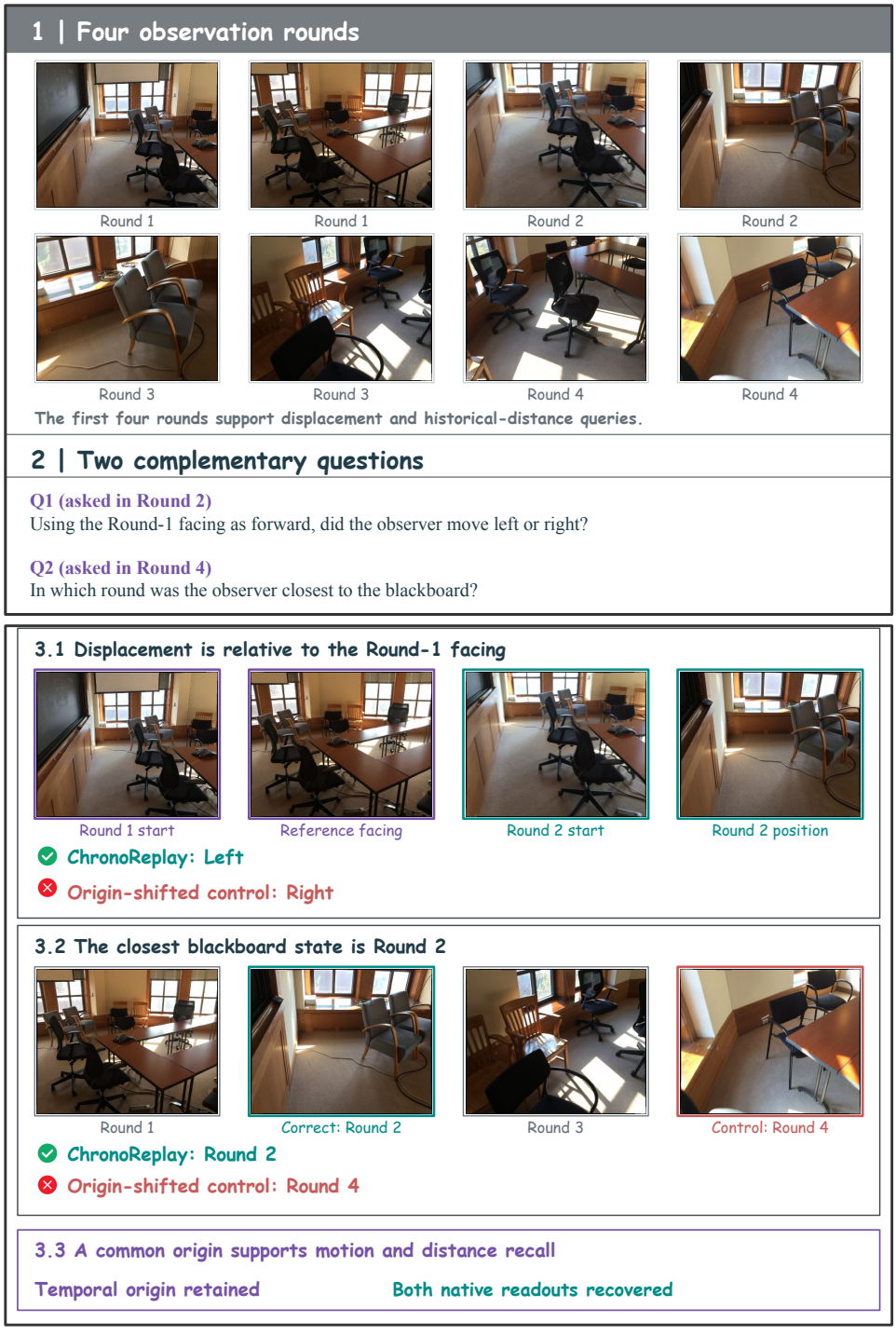


Figure 7: **Reference displacement and nearest-state retrieval.** In scene `scene0496_00`, ChronoReplay identifies leftward displacement relative to the Round 1 facing, while the control answers right. When asked in Round 4 for the closest blackboard state, ChronoReplay retrieves Round 2 rather than the control's Round 4. Left/right denotes displacement, not turning direction; the closest-state label follows the source annotation rather than apparent object size in a single image.

F.4 Native Numerical Estimation and Supervision Complementarity

Figure 8 compares Base, historical Preservation and Trajectory diagnostic models, and Joint replay on two questions about the same ornate mirror in OST-Bench scene `1mp3d_0004_region16`. The numerical question is asked at the end of Round 3 and the categorical question at the end of Round 4; this is a shared episode with different query endpoints, not two questions conditioned on an identical complete history. All four settings receive the same source-ordered images and question prompts. The historical Preservation and Trajectory models differ from the single-family variants in Table 4(a). As in the official multi-round protocol, each model retains its own earlier responses. The Round 4 observations are not exposed to the Round 3 query.

Different strengths, rather than a uniform winner. The source annotation places the mirror on the observer's right in Round 4. Base, Preservation, and Joint replay answer right, whereas Trajectory answers left. The Round 3 horizontal distance is annotated as 0.95 m; the four predictions are 1.20, 2.00, 1.10, and 1.11 m, respectively. Trajectory has the smallest absolute error, 0.15 m, followed by Joint at 0.16 m. Thus Joint retains the correct categorical answer while approaching the numerical accuracy of Trajectory; it is not best on every individual readout.

Official error convention. For this nonzero scalar target y, OST uses relative error $\delta = |\hat{y} - y|/|y|$ and awards the fraction of ten strict thresholds satisfied:

$$s_{\text{est}} = \frac{1}{10} \sum_{i=0}^{9} \mathbf{1}[\delta < 0.50 - 0.05i] \,. \tag{15}$$

The respective relative errors are 26.3%, 110.5%, 15.8%, and 16.8%, giving official scores 0.50, 0.00, 0.70, and 0.70. Trajectory and Joint tie under this discretized metric despite their different raw errors. Distances and directions follow the source annotation, not measurements inferred from RGB appearance. This selected episode illustrates the aggregate trade-off; it does not estimate its frequency.

F.5 A Global Layout Regression under Fixed Observations

Figure 9 shows OVO-S L4.3 record `4.3_034`, an ARKitScenes room tour ending at 50 seconds. The question asks which supplied bird's-eye view matches the scene. Under the same Uniform-128 observations, the official Base selects the annotated option B, while the headline ChronoReplay checkpoint selects D. Both archived records use the same default prompt, ordered image blocks, greedy categorical generation, and unpermuted option labels. This pair isolates an output regression under fixed access; it is not an adaptive-versus-uniform comparison.

The concrete disagreement. The video reveals the fireplace, television, table, and surrounding furniture across successive views. In the shared coordinate frame of the supplied option diagrams, B places the fireplace and TV in the lower region and the table across the layout. D instead places the fireplace in the upper region and moves the table beside the TV in the lower-right region. The figure pairs key video observations with the original B and D option diagrams, labelled by the model that selected each option. All four layouts are retained in the supplementary source assets. The source provides no evidence-time intervals for this item, so the displayed frames are selected from the recorded sampling protocol, not labelled as official evidence spans. This discrepancy demonstrates a global-layout failure. It does not establish that the model built and then lost a global map, nor isolate perception, cross-view integration, or option matching as the internal cause.

Exact questions, model paths, responses, numerical scores, frame references, and all four source layout options accompany the manuscript in `supplementary/spatial_capability_cases.json`. No new model inference is used for either case.

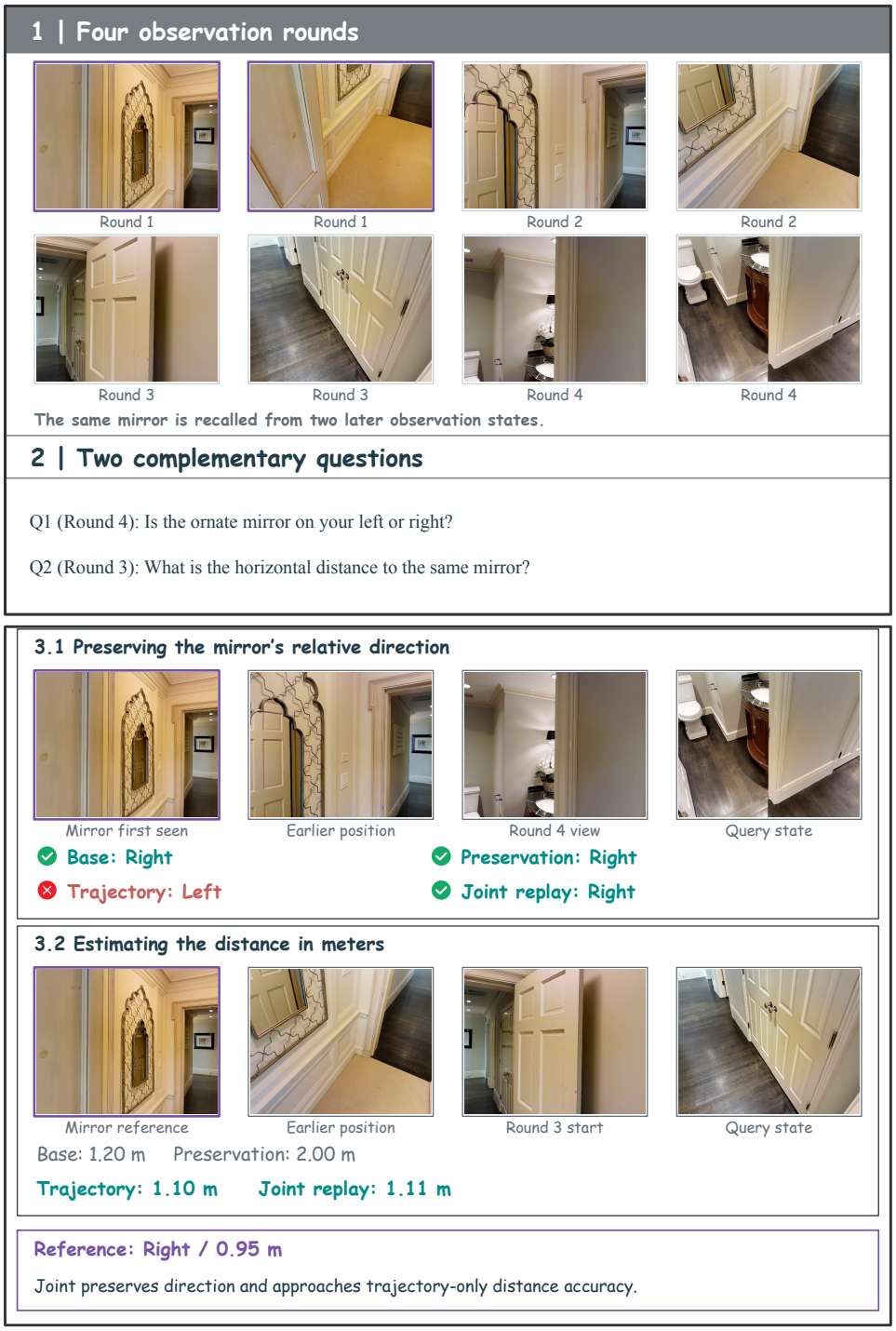
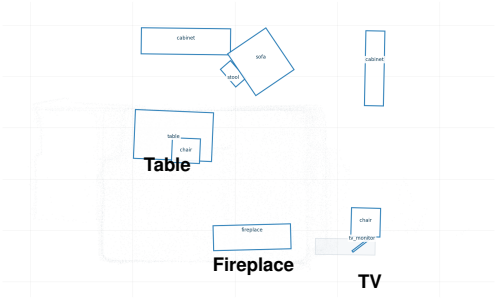


Figure 8: **Two native interfaces on one mirror history.** Selected chronological observations precede a left/right judgement and a horizontal-distance estimate. Round labels identify the source observation turns. Base, Preservation, Trajectory, and Joint replay use the same observations at each query endpoint. Joint retains the correct direction; Trajectory has slightly lower absolute distance error, and both tie at 0.70 under the official metric.

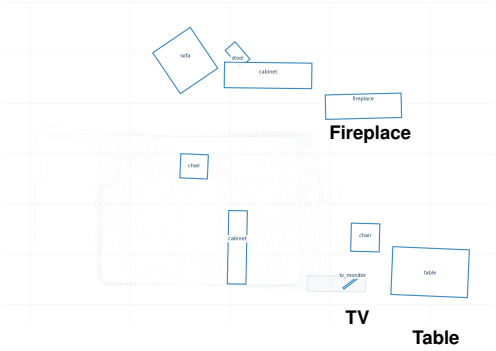


B: annotated layout / Base



Fireplace and TV below the table

D: ChronoReplay's choice



Fireplace above TV; table right of TV

Figure 9: **A fixed-observation global-layout failure.** The official Base chooses B and ChronoReplay chooses D on OVO-S L4.3. The upper panel shows eight timestamped frames from their shared Uniform-128 input. The lower panels reproduce the two source option diagrams in their original orientations, with enlarged labels beside the key objects. Captions describe positions within these diagrams. The fireplace and table placements make the disagreement visible.

G SUPPLEMENTARY BENCHMARK RESULTS

G.1 OVO-S-BENCH

Table 1 reports the complete twelve-family decomposition for the official base and ChronoReplay under Uniform-128 and adaptive access. Table 5 reports the corresponding level means. The independent Uniform-81 control is in Appendix E.3. The supplementary table below provides additional published aggregate reference scores; the evaluation population and comparison scope are defined in Appendix B.

Table 16: Additional published OVO-S aggregate reference rows (%). These rows provide external context and are not used for matched training attribution.

Model	Params	L1	L2	L3	L4	Overall
Proprietary MLLMs						
GPT-5.4	–	54.64	57.59	50.80	40.53	50.89
Gemini-3.1-Pro	–	61.92	64.04	55.90	54.90	59.19
Gemini-3.1-Flash-Lite	–	54.09	52.19	54.14	42.84	50.81
Grok-4.1-Fast	–	44.80	46.59	48.49	35.04	43.73
Open-source MLLMs						
InternVL-3.5	38B	54.65	54.40	45.53	41.94	49.13
InternVL-3.5	241B-A28B	55.55	55.72	51.62	40.54	50.85
Qwen3-VL	4B	43.25	48.18	54.54	41.35	46.83
Qwen3-VL	32B	50.14	51.91	51.94	41.22	48.80
Qwen3-VL	235B-A22B	52.45	55.18	61.20	45.71	53.64

G.2 OST-BENCH

Table 17 retains the aggregate OST summary and the complete published reference set used in this paper. The same native scorer and scene-disjoint manifest are used for the official base and ChronoReplay rows.

Table 17: Aggregate OST-Bench summary and published reference set (%). Published rows use the released full split; the final block is compared only within the matched scene-disjoint protocol.

Model	Overall	A. State	A. Info	AO	Judgement	Temporal	Counting	Estimation
Proprietary MLLMs (published)								
Claude-3.5-Sonnet	47.8	45.6	65.6	32.9	57.4	46.7	57.9	26.9
Gemini-2.0-Flash	49.5	45.1	70.8	33.8	61.1	45.8	59.4	28.1
Gemini-2.0-Flash (Thinking)	54.2	47.1	72.3	42.8	63.6	61.1	61.9	28.9
GPT-4o	48.7	38.8	72.8	33.5	59.8	48.6	59.8	24.0
GPT-4.1	53.4	47.2	76.5	37.7	66.4	51.7	60.6	29.1
Open-source MLLMs (published)								
InternVL-2.5-8B	44.6	41.9	58.6	34.6	52.4	42.7	56.4	29.3
InternVL-2.5-38B	50.8	45.4	73.9	34.0	61.4	45.9	61.1	31.5
InternVL-2.5-78B	51.1	46.5	74.0	32.9	61.5	47.1	65.9	29.0
Qwen2.5-VL-7B	41.2	40.4	52.6	31.5	50.1	36.6	62.1	22.7
Qwen2.5-VL-72B	45.6	43.5	64.5	28.9	55.9	41.6	61.5	22.9
Matched scene-disjoint evaluation								
Official base	50.52	48.28	68.96	36.36	60.70	44.86	61.37	33.80
ChronoReplay	63.93	53.83	84.69	51.78	72.72	62.45	78.54	43.63

The gain spans every reported component, including continuous Estimation and Temporal-localization. The matched origin experiment is reported separately in the temporal-origin block of Table 4; it does not replace this checkpoint-level comparison.

G.3 UCS-BENCH

Table 3 reports four category aggregates and eight UCS subcategories over the complete 8,114-question evaluation. The table below adds the aggregate published reference scores and matched

model comparison. Appendix B defines causal access and scoring scope. Training-topology robustness is evaluated separately in Appendix D.6.

Table 18: Published UCS aggregate reference set and matched 8,114-question evaluation (%). Published rows use their reported frame protocols.

Model	Size	Frames	Trajectory	Position	Proximity	Category	Overall
Human and proprietary references							
Human	–	–	91.7	87.6	96.8	90.0	91.0
DeepSeek-V3	–	50	42.0	33.2	37.7	32.3	35.6
Doubao-Seed-1.6V	–	50	49.6	42.8	56.0	44.6	47.3
GPT-5	–	50	44.5	37.8	55.4	41.4	43.7
Gemini-3-Flash	–	50	43.8	38.0	52.9	38.9	42.2
Open-source MLLMs							
InternVideo2.5	8B	64	44.2	39.4	52.7	42.2	43.8
Qwen2.5-VL	7B	64	41.6	40.8	51.3	42.5	43.4
Qwen2.5-VL	32B	64	50.3	40.9	55.4	46.0	47.3
Qwen3-VL	8B	64	44.6	40.1	50.9	42.9	44.0
Qwen3-VL	32B	64	50.3	46.5	55.2	46.7	49.0
DirectMe (Qwen3-VL)	8B	1fps	54.1	47.6	58.0	46.4	50.5
Matched evaluation							
Official Qwen3-VL-8B	8B	128	48.49	41.72	49.49	43.87	45.33
ChronoReplay	8B	128	56.14	48.26	54.35	53.17	52.77

H Relation-Supervision Diagnostics

This appendix reports a negative diagnostic: whether adding a more discriminative auxiliary relation interface, a binary compatibility judgement between a causal state and a typed relation claim, yields a more reusable native representation. These variants do not contribute to the headline ChronoReplay checkpoint or its training data. Relation rows are compiled from verified source relations (signed heading, metric displacement, temporal retrieval, counting, topology, and local state) paired with controlled incompatible alternatives that stay in the same value domain; positive and incompatible claims are independent continuations of one observation prefix, and only the final compatibility token receives loss. Source question, candidate order, option letter, benchmark identity, and measured score are absent from the state prefix. Within-state normalization assigns fixed mass to each causal state regardless of the number of relations K; the local $K = 1$ control adds 1,920 rows over 960 states, and cross-type $K = 4$ adds 7,680 rows over 960 states.

Table 19: Overview of relation-supervision diagnostics. Native columns are complete model-only benchmark scores; Readout is the matched withheld compatibility diagnostic (%).

Variant	OVO-S	OST	UCS	Readout
ChronoReplay, no relation loss	53.89	63.93	52.77	–
Local compatibility, $K = 1$	53.73	64.08	52.65	63.33
Same type, $K = 4$	53.50	64.06	51.70	62.29
Cross type, $K = 4$	52.07	64.34	51.77	66.88
State-path firewall	52.18	63.42	54.41	58.13

Cross-type supervision improves the withheld relation readout over the matched same-type control, but its native three-endpoint mean is lower. The state-path firewall, which restricts relation gradients to the visual-state route, raises UCS while reducing OVO-S and OST. Formulation controls on complete OST and the 480-row withheld readout behave the same way: removing explicit type names gives 64.45 OST and 66.67 readout, presenting candidates jointly gives 64.57 and 63.96, and removing state normalization gives 64.54 and 67.92, so each moves the local readout without improving native transfer. An unbalanced-polarity variant reaches 97.08% on incompatible claims but only 28.75% on compatible ones (64.34 OST). Auxiliary separability is therefore not sufficient for native cross-interface transfer, which the main paper evaluates through the native benchmarks.

I Reproducibility Details

I.1 Executable Configuration

All runs start from the pinned `Qwen/Qwen3-VL-8B-Instruct` revision listed in Table 20 and reuse its tokenizer, processor, and chat template. Training follows Table 8, without adapters, continuation, fusion, curriculum, RL, or offload.

Table 20: Execution settings for the headline, target-format, and chronology comparisons in Appendix D.

Setting	Value
Base-model revision	`0c351dd01ed87e9c1b53cbc748cba10e6187ff3b`
Headline / target-format seed	42
Chronology seeds	42 (Run 1); 73 (Run 2)
Headline and chronology topology	8 GPUs; gradient accumulation 2
Target-interface topology	4 GPUs; gradient accumulation 4

Inference is deterministic. OVO-S uses greedy answer generation with the fixed evaluation prompt and deterministic answer extraction, an eight-token ceiling, and batch size two. OST uses greedy decoding with temperature 0, top-$p = 1$, one sample, and 512 tokens; UCS uses temperature 0, one sample, and eight tokens. No voting, prompt ensemble, or benchmark-dependent parameter choice contributes to a paper row.

I.2 Data and Media Audits

All training examples are bound to canonical scene identities and source annotations before optimization. The same data separation is applied to the main model and every matched control. Complete media, row, and scene manifests are retained with the released configuration.

Each release binds row hashes, counts, family mass, supervised turns, and the media manifest. Every referenced image and video is decoded and audited once before training: items at or above the 32,768-token limit or above the declared visual-pixel budget are rejected, and isolation audits bind canonical physical scenes.

I.3 Scorer and Manifest Contracts

Each reported result binds the model, observations, predictions, annotations, and scoring protocol. Evaluation follows the fixed OVO-S protocol in Appendix B, the fixed 8,351-turn OST manifest, and all 8,114 UCS questions. Result records and source manifests link every evaluation identity to the values defined in `main.tex` and will be released with the code.

Native responses are passed to the unchanged task scorer. For canonical responses, a frozen reader extracts the leading typed `value` field from the generated response alone and supplies that value to the same scorer; it does not inspect the question, task identity, reference answer, label, or score. All 8,351 OST responses and all 8,114 UCS responses yield an extracted value. The official OST scorer evaluates its designated 7,814 turns.