EVENTTIMEQA: AUDITING INTERVENTION-CONSISTENT MILLISECOND TEMPORAL REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Always choosing the same side yields 50% sample accuracy on a factual/counterfactual pair while never answering a pair completely correctly. EventTimeQA uses this distinction to test millisecond-scale temporal perception in controlled synthetic scenes. Across 10,000 pairs, exactly one of onset order, relative flicker frequency, or time to contact changes while all scene nuisance variables stay fixed. We report pair accuracy and prediction flips under changes in temporal resolution, timestamps, scene and sensor settings, and post-simulation event retention. Lightweight supervised event models approach ceiling performance on the frozen v3 dataset. For vision-language models, the input interface sharply changes the outcome: most models remain weak on event contact-sheet input, whereas a budget-compliant externally evaluated Gemini model reaches 74.0% pair accuracy on RGB500 and at most 31.9% on event input. The leakage-audited 0/2/4/8-shot results vary with model and interface and are not monotonic in shot count. The release includes construction and statistical protocols, shortcut checks, documentation, and maintenance procedures. All evidence is confined to synthetic scenes; physical-sensor transfer and human performance remain untested.

1 INTRODUCTION

An isolated correct answer says little about which part of a temporal sequence the model actually used. A predictor that ignores timing can still benefit from nuisance correlations; on a balanced pair, always choosing the same side already gets one of the two answers right.

The central test changes one millisecond-scale variable and reverses the label while leaving the two examples indistinguishable at low frame rates. A model that perceives the change should reverse its prediction with the label.

Event cameras record per-pixel brightness changes asynchronously, with high temporal resolution and dynamic range (Lichtsteiner et al., 2008; Gallego et al., 2022). The field has produced datasets converted from static images (Orchard et al., 2015), driving recordings (Gehrig et al., 2021), and methods for reconstructing frames from events (Rebecq et al., 2019). EventGPT and EventBench span broader semantic tasks (Liu et al., 2025b;a). For video, FPS-Bench and TemporalBench probe fine temporal understanding (Choudhury et al., 2026; Cai et al., 2024); Vinoground, CLEVRER, and CaST-Bench address temporal counterfactuals or physical causality (Zhang et al., 2024; Yi et al., 2019; Zhang et al., 2026). Work on event representations ranges from hand-crafted time-surface histograms (Sironi et al., 2018) and differentiable recurrent surfaces (Cannici et al., 2020) to point-cloud encoders (Ren et al., 2024), open-world representations (Zhou et al., 2024), and adversarial event generation (Ren et al., 2026a); see also the survey by Ren et al. (2026b). Our question is narrower: does a model follow a registered change in a single temporal variable when all other scene factors stay fixed?

We organize the audit around three research questions. **RQ1 (construct isolation):** Does the temporal signal survive shortcut controls, different event-retention policies, and a separately implemented generator? **RQ2 (capability mapping):** How do results vary with temporal resolution, representation, model family, interface, and in-context examples under a common evidence budget? **RQ3 (controlled**

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.

Benchmark	Events	Paired CF	Single-var.
EventGPT	✓	–	–
EventBench	✓	–	–
FPS-Bench	–	–	–
TemporalBench	–	–	–
Vinoground	–	✓	–
CLEVRER	–	questions	–
CaST-Bench	–	questions	–
EventTimeQA	✓	✓	✓

Table 1: Comparison with event and video benchmarks. "Paired CF" distinguishes paired counterfactual examples from counterfactual questions; "Single-var." denotes a registered single temporal-variable swap. EventTimeQA additionally reports a timestamp audit. A dash means the feature was not reported in the versions reviewed here; it does not establish that the feature is absent from all materials.

robustness): Which findings survive shifts in scene style, sensor settings, retention policy, and synthetic generator?

The release includes paired temporal interventions, a controlled generator with three audited event-retention branches, and evaluations across representations, resolutions, and synthetic domains. Pair-cluster statistical inference accompanies all reported results. We also provide a leakage-audited multimodal 0/2/4/8-shot protocol; its selection and scoring details appear in Appendix A.

2 Task and Positioning

The benchmark tests consistency under a millisecond-scale temporal intervention. Each sample presents two objects or stimuli and asks which started moving first, which flickers faster, or which will reach a target first. Within a factual/counterfactual pair, the temporal variable that determines the answer changes while all other visual elements remain fixed. A pair counts as correct only if both answers are correct. Table 1 compares this task with related event and video benchmarks.

The comparison concerns the intervention and evidence, rather than question breadth. EventGPT and EventBench cover event semantics; FPS-Bench and TemporalBench examine fine timing in video. Vinoground includes paired temporal counterfactuals, while CLEVRER and CaST-Bench ask causal questions. In the reviewed versions, none reports the same combination of a registered single-variable temporal swap, event input, and an explicit timestamp audit. The qualified dashes in Table 1 should not be read as claims that these benchmarks cannot support temporal or causal evaluation.

3 Benchmark Construction

Each scene is rendered in grayscale at 500 FPS, then passed to a revision-pinned v2e emulator (Hu et al., 2021). As with ESIM (Rebecq et al., 2018), v2e turns intensity changes into asynchronous events. Within a pair, background, appearance, question, final state, nuisance values, and sensor parameters are shared between the two members. Only the registered temporal parameter changes, and the answer changes with it. We split at the pair level into in-distribution (ID), held-out scene-style, and held-out sensor groups.

The reported experiments use a fixed set of 10,000 pairs (frozen v3). After event simulation, deterministic first-retained thinning equalizes the 32 ms (bin, x, y, p) multisets. The frozen-v3 streams are therefore processed rather than raw, and we label them as such. Figure 1 places the frozen policy beside the unmodified and randomly retained control branches.

Scope of the matching intervention. Matching applies to a coarse 32 ms signature, not to the raw event-time sequence. Finer timing information remains available to answer the question. A representation that retains only the matched 32 ms multiset receives identical evidence from the two members. With the question also fixed, a deterministic predictor operating on that representation must return the same answer twice and cannot make the pair correct. This explains why the coarse-bin condition is an information-removal control: failure there can arise before model inference. Conversely,

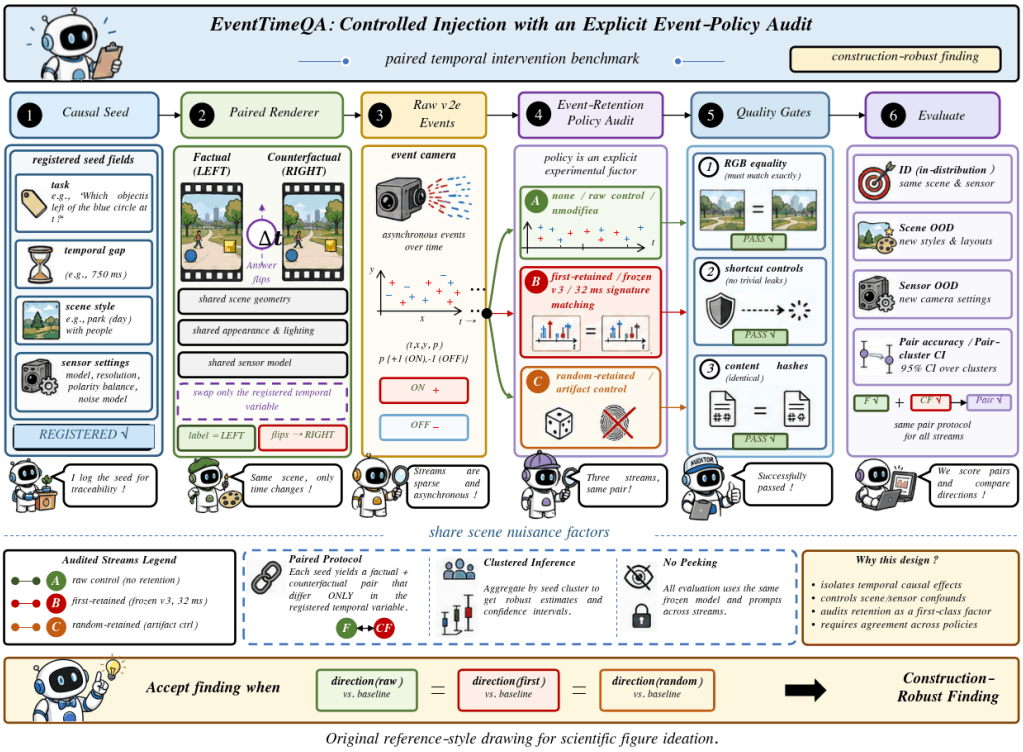


Figure 1: **Construction and retention-policy audit.** Causal seeds produce paired renders and raw v2e streams. The unmodified, frozen-v3 (first-retained), and deterministic random-retained streams all undergo the same shortcut, integrity, and pair-level checks.

successful prediction from finer bins shows that the released input retains useful distinctions, without establishing how the model internally represents them. First-retained thinning occurs after simulation and belongs to the released representation, not to the event generator. A model tested only on frozen v3 might therefore exploit either the intended timing or a trace of the retention rule. We check that possibility with unmodified streams and a random-retained branch; both controls remain synthetic.

In the 1,000-pair-per-policy audit (Table 11), 27.7% of unmodified v2e pairs already have equal 32 ms signatures. First and random retention keep 94.72% of events on average, with a per-pair minimum of 56.32%. A bilateral ridge classifier given event-time histograms but no spatial coordinates serves as a temporal positive control. The unmatched streams also carry a mild nuisance signal. The unmodified temporal control has a 95% pair-cluster confidence interval of $[88.62, 96.41]\%$. This interval describes uncertainty in one within-policy estimate; it does not measure transfer between policies.

Architecture choice matters more in the $3 \times 3 \times 3$ transfer matrix (architecture $\times$ policy $\times$ seed). GRU ID pair accuracy ranges from 85.2% to 91.0%, whereas the other architectures are unstable or near chance under this training budget. One-sided temporal inputs alone reach 82.0–86.8% within policy. Timestamps are useful, but a model need not compare both sides explicitly to do well. We consequently interpret the task as temporal-intervention classification rather than a test of compositional reasoning.

Shortcut controls. Raw and retained streams must be reported separately: on unmodified events the count-only and no-time ridge controls reach 55.99% and 56.89% ID sample accuracy (Table 11), above the 50% of the metadata control. After either retention rule, all three nuisance-only controls sit at exactly 50%. The bilateral temporal-histogram control, by contrast, scores 92.81%, 97.01%, and 95.81% pair accuracy on unmodified, first-retained, and random-retained streams (Figure 2). Unmatched streams thus show above-chance scores on these nuisance controls; matching removes that

residual leak without removing the fine temporal signal. Other shortcuts remain possible, motivating the alternative-retention controls and separate-generator evaluation.

To test dependence on v2e, we built a separate contrast-threshold simulator that shares no event-generation code with v2e; it keeps a per-pixel log-intensity reference and interpolates threshold crossings. We generated 1,000 new pairs in both unmodified and frozen-v3-matched forms. Source-trained GRU checkpoints reach 100% pair accuracy on both forms. Transformer and Temporal Conv checkpoints trained on all 10,000 source pairs reach 99.97% and approximately 97.3%. Agreement between two synthetic generators does not establish transfer to physical sensors.

4 Evaluation Protocol

A pair counts as correct only when its factual and counterfactual answers are both correct. We additionally report sample accuracy, flip consistency, per-task and temporal-gap results, timestamp-shuffle effects, 2–64 ms event-bin curves, 30–500 FPS controls, parse failures, and ID-to-out-of-distribution (OOD) changes. The model groups are supervised event encoders, pretrained event models, general vision-language models (VLMs) evaluated with zero or few examples, and benchmark-trained adaptation diagnostics. Every multimodal media item uses the registered evidence window with exactly 12 temporal units.

Why the pair is the unit of analysis. The gold labels are opposite within each pair. Predicting the same side twice gives one correct sample and one incorrect sample: 50% sample accuracy, zero pair accuracy. A model can also flip its answer in the wrong direction, earning flip consistency while missing both labels. We report all three quantities to keep these behaviors distinct. For the same reason, uncertainty is computed over pair IDs. Resampling the two members separately would discard the dependence introduced by construction.

More formally, let c_i^{f} and c_i^{c} indicate whether the two answers in pair i are correct. Sample accuracy averages $(c_i^{\mathrm{f}} + c_i^{\mathrm{c}})/2$, whereas pair accuracy averages $c_i^{\mathrm{f}} c_i^{\mathrm{c}}$. The product rejects a correct answer that is not supported by a correct response to the intervention. For valid binary answers, a prediction flip is either correct on both members or inverted on both; a constant prediction is correct on exactly one. Thus, high flip consistency alone need not indicate useful temporal perception. Independent uniform guesses have expected pair accuracy 25%, while constant-side predictions have pair accuracy zero. These are different reference behaviors and must not share a chance label. Parse failures remain failures under the registered scoring rule.

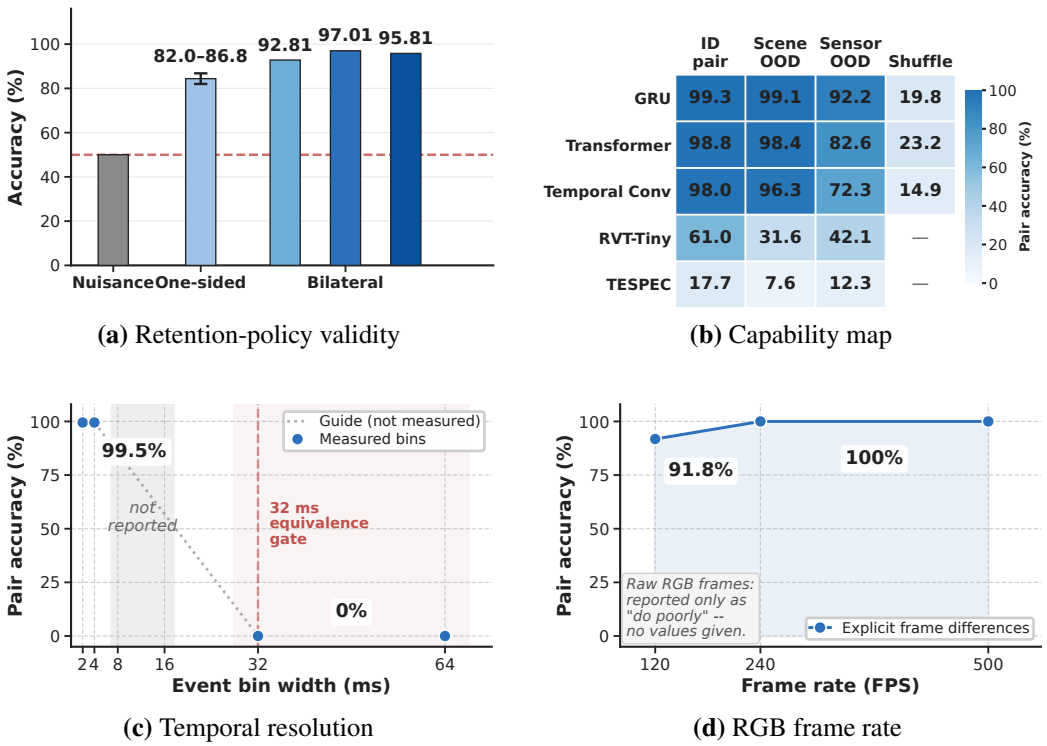
For few-shot evaluation, $k \in \{0, 2, 4, 8\}$ is the number of individual demonstrations, drawn as $k/2$ complete factual/counterfactual training pairs. Selection is deterministic, nested across shot counts, balanced by answer side, and based only on the visible task. An overlap check rejects any demonstration that shares a pair ID, sample ID, or media file with its query. Each query and demonstration media item carries 12 temporal units, so the total context contains $12(k + 1)$ units. Every model–interface–shot-count combination uses the same 1,000 complete ID test pairs.

Adding demonstrations lengthens the context while preserving query evidence. Higher-shot prompts extend lower-shot prompts with complete training pairs; the overlap audit excludes the query's counterfactual counterpart. Appendix A gives the selection conditions and metric cells.

Confidence intervals use 10,000 bootstrap resamples of complete pairs. Paired model comparisons use pair-cluster bootstrap and sign-flip permutation tests; Holm family-wise and Benjamini–Hochberg false-discovery corrections are reported within each declared family.

5 Controlled-Synthetic Findings

Retention-policy controls and generator transfer. The bilateral temporal control remains accurate under all three within-policy evaluations, scoring 92.81–97.01%. The available summaries do not provide a complete train-policy by test-policy transfer matrix, so they do not establish accuracy in every transfer direction. On the independent contrast-threshold generator, source-trained GRU checkpoints reach 100% with and without policy matching; full-data Transformer and Temporal Conv checkpoints reach 99.97% and approximately 97.3%. The independent-generator results are inconsistent with an explanation that requires the original simulator and its first-retained output alone.



(a) Retention-policy validity

(b) Capability map

(c) Temporal resolution

(d) RGB frame rate

Figure 2: **Key experimental results.** (a) Retention-policy validity: after retention, the three nuisance-only ridge controls sit at chance (50.0%), one-sided temporal inputs reach 82.0–86.8% (bar shows the reported range), and the bilateral temporal-histogram control scores 92.81%, 97.01%, and 95.81% on unmodified, first-retained, and random-retained streams. (b) Capability map: supervised encoders cluster near ceiling on ID and scene-OOD pairs, separate under sensor shift, and collapse when event–time correspondence is shuffled; frozen backbones sit far lower, and no shuffle value is reported for them. (c) Temporal resolution: pair accuracy is ≈99.5% at 2–4 ms bins and 0% at 32–64 ms; the 8 and 16 ms bins are shaded because they are not reported here, and the dotted segment is a visual guide, not a measurement. (d) RGB frame rate: explicit frame differences reach 91.8% at 120 FPS and 100% at 240 and 500 FPS; raw RGB is described only qualitatively in the text, so no curve is drawn for it.

Still, one-sided temporal inputs score 82.0–86.8% within policy: the evidence supports classification from temporal information, not the stronger claim that bilateral compositional reasoning is required.

Temporal resolution and RGB differencing. Pair accuracy on binned events is approximately 99.5% at 2–4 ms and zero at 32–64 ms. Raw RGB frames perform poorly at every frame rate tested. Supplying explicit frame differences to the same architecture reverses this: pair accuracy rises to 91.8% at 120 FPS and 100% at 240 and 500 FPS. The construction therefore demands fine temporal evidence, but under this protocol event input shows no intrinsic advantage over suitably processed high-frame-rate RGB (Figure 2).

Sample accuracy hides pair failures. The three supervised sequence architectures separate under sensor shift (Table 2), although each loses consistency when event–time correspondence is shuffled. Among the supervised models tested here, the continuous event GRU reaches 99.33% ID pair accuracy. The frozen backbones show a different regime: RVT-Tiny pairs 80.07% sample accuracy with 61.00% pair accuracy, and TESPEC 58.13% with 17.67% (Table 3). The best open-VLM event-contact-sheet interface tested reaches 13.0% pair accuracy. These figures describe the representations and interfaces tested here; they do not establish a ceiling for any entire model family.

The near-ceiling ID column compresses differences among the three trained encoders. Scene OOD remains above 96% for all three, while sensor OOD ranges from 92.20% (GRU) to 72.27% (Temporal Conv). The latter also has a large three-seed standard deviation (11.93 points), so a precise ranking

Supervised event model	Params.	ID pair	Shuffled pair	Scene OOD	Sensor OOD
GRU	119K	99.33 ± 0.06	19.83 ± 5.44	99.13 ± 0.12	92.20 ± 2.03
Transformer	631K	98.80 ± 0.17	23.23 ± 0.87	98.40 ± 0.35	82.60 ± 2.65
Temporal Conv	353K	98.00 ± 2.17	14.87 ± 4.51	96.33 ± 4.45	72.27 ± 11.93

Table 2: **Supervised event architectures.** Pair accuracy in percent, mean $\pm$ standard deviation over three seeds. Training data, event budget, evidence window, and test pairs are fixed across rows. Sensor OOD separates the models more clearly than the near-ceiling ID results do.

Frozen event backbone	ID sample	ID pair	Scene OOD pair	Sensor OOD pair
TESPEC (Mohammadi et al., 2025)	58.13 ± 1.26	17.67 ± 2.75	7.60 ± 1.20	12.27 ± 2.31
RVT (Gehrig & Scaramuzza, 2023)	80.07 ± 0.19	61.00 ± 0.46	31.60 ± 0.92	42.07 ± 0.58

Table 3: **Frozen pretrained event backbones.** Values are percentages, mean $\pm$ standard deviation over three probe-training seeds. The input window, representation, probe, labels, and splits are fixed.

should not be inferred from its mean alone. Shuffling event–time correspondence lowers pair accuracy to 14.87–23.23% across architectures, well below their unshuffled ID values. Taken together, these results show that the tested encoders rely on timing, and that sensor variation separates them more clearly than the scene-style shift does. They do not identify any particular sensor parameter as the cause of the gap.

We pin the official TESPEC (Mohammadi et al., 2025) and RVT (Gehrig & Scaramuzza, 2023) revisions and checkpoints, freeze every backbone parameter, construct their native 20-channel stacked histograms over the same registered evidence window, and train an identical task-aware probe for each.

Freezing the backbones makes this a probe comparison rather than full task fine-tuning. Despite identical input construction and probes, RVT-Tiny drops from 61.00% ID pair accuracy to 31.60% on scene OOD and 42.07% on sensor OOD; TESPEC drops from 17.67% to 7.60% and 12.27%. Different pretraining histories prevent an architecture-only interpretation. The supervised encoders instead learn from benchmark labels, so their higher scores answer a different question. Figure 2 summarizes these evaluation conditions.

Few-shot changes are not monotonic. All 16 Qwen cells in Table 4 follow the strict protocol, as does Gemma-3-12B-it. With Qwen3-VL-2B on event input, two demonstrations raise pair accuracy from 0.3% to 4.8% ($+4.5$ points, pair-cluster 95% CI $[3.2, 5.9]$, Holm-adjusted $p = 0.0009$). Four and eight demonstrations bring it back down to 1.3% and 0.1%, respectively. On native RGB500, eight demonstrations raise accuracy from 1.3% to 5.3% ($+4.0$ points, 95% CI $[2.5, 5.6]$, Holm-adjusted $p = 0.0009$); four reduce it by 1.1 points (95% CI $[-1.9, -0.4]$, Holm-adjusted $p = 0.0468$). Qwen2.5-VL-7B remains at 0–0.1% pair accuracy and predicts a constant side on at least 99.9% of pairs in every cell.

Gemma-3-12B-it differs across inputs: its event-contact-sheet pair accuracy rises from 1.4% at zero shots to 5.0% at four, while RGB500 falls from 2.7% to 0.7% at eight. Both changes have Holm-adjusted $p < 0.05$. Demonstrations therefore do not reliably help: the effect depends on model and interface. Appendix A reports the remaining statistics.

In the supplementary evaluation, Gemini-3.1-Flash-Lite reaches 25.9%, 31.9%, 31.8%, and 30.4% event pair accuracy at 0, 2, 4, and 8 shots. The corresponding RGB500 values are 68.0%, 74.0%, 68.0%, and 72.7%. All eight cells contain 2,000 samples with zero parse failures and satisfy the requested eight-token output and disabled-reasoning controls. Its upstream revision and direct-provider provenance are unverified.

The Qwen checkpoints are revision-pinned. All 16 Qwen cells contain 2,000 samples (1,000 complete pairs), with no parse failures and the same nested training demonstrations. At the evidence freeze, EventVL (Li et al., 2025) and EventFlash (Liu et al., 2026) were not evaluated because compatible checkpoints were unavailable to this evaluation, while EventDrive (Lu et al., 2026) required a task adapter outside the registered protocol. None of the three is counted as a completed row. We report the frozen soft-prefix projector only as an interface diagnostic because its event encoder was trained

Model	Input	0-shot	2-shot	4-shot	8-shot
Qwen3-VL-2B	event	0.3 [0.0,0.7]	**4.8 [3.5,6.2]**	1.3 [0.6,2.0]	0.1 [0.0,0.3]
Qwen3-VL-2B	rgb500	1.3 [0.6,2.0]	0.8 [0.3,1.4]	0.2 [0.0,0.5]	**5.3 [3.9,6.7]**
Qwen2.5-VL-7B	event	0.0 [0.0,0.0]	0.0 [0.0,0.0]	0.0 [0.0,0.0]	0.0 [0.0,0.0]
Qwen2.5-VL-7B	rgb500	**0.1 [0.0,0.3]**	0.0 [0.0,0.0]	0.0 [0.0,0.0]	0.0 [0.0,0.0]
Gemma-3-12B-it	event	1.4 [0.7,2.2]	4.3 [3.1,5.6]	**5.0 [3.7,6.4]**	4.7 [3.4,6.0]
Gemma-3-12B-it	rgb500	2.7 [1.7,3.8]	1.0 [0.4,1.7]	0.8 [0.3,1.4]	0.7 [0.2,1.3]

Table 4: **Pair-isolated multimodal in-context learning.** Entries are ID pair accuracy in percent with pair-cluster 95% confidence intervals over 1,000 complete test pairs. Each shot is one multimodal example; demonstrations comprise complete training pairs, and every media item contains exactly 12 temporal units. Bold marks the best shot count within each model-interface row.

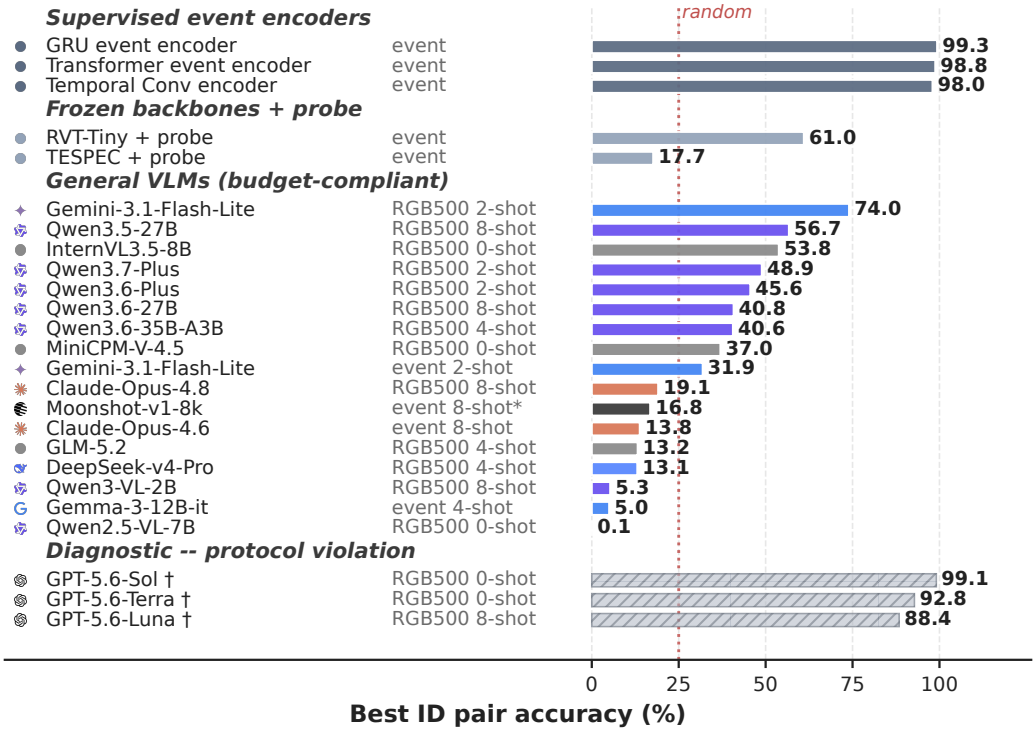


Figure 3: **Pair accuracy across model groups.** Each bar is the best ID pair accuracy for the indicated model, interface, and shot count, from Tables 2, 3, 10, 7, and 8–9. Groups distinguish supervised encoders, frozen probes, budget-compliant VLMs, and diagnostic outputs. The dotted line is the 25% expectation for independent uniform binary guesses; a constant-side predictor has zero pair accuracy. †GPT-5.6 evaluations violate the output or reasoning controls and are diagnostic only. *Moonshot's 8-shot condition packs demonstrations into four images and changes the presentation format. External evaluations lack verified upstream revisions. Different training histories and interfaces preclude a controlled cross-group ranking. Vendor logos identify the reported model families; sources are `simple-icons` and `gilbarbara/logos`.

on EventTimeQA labels. Figure 3 collects the budget-compliant pair-accuracy rows across the model groups; representations and training histories differ across groups, so the rows are grouped by family and should not be read as a single ranking.

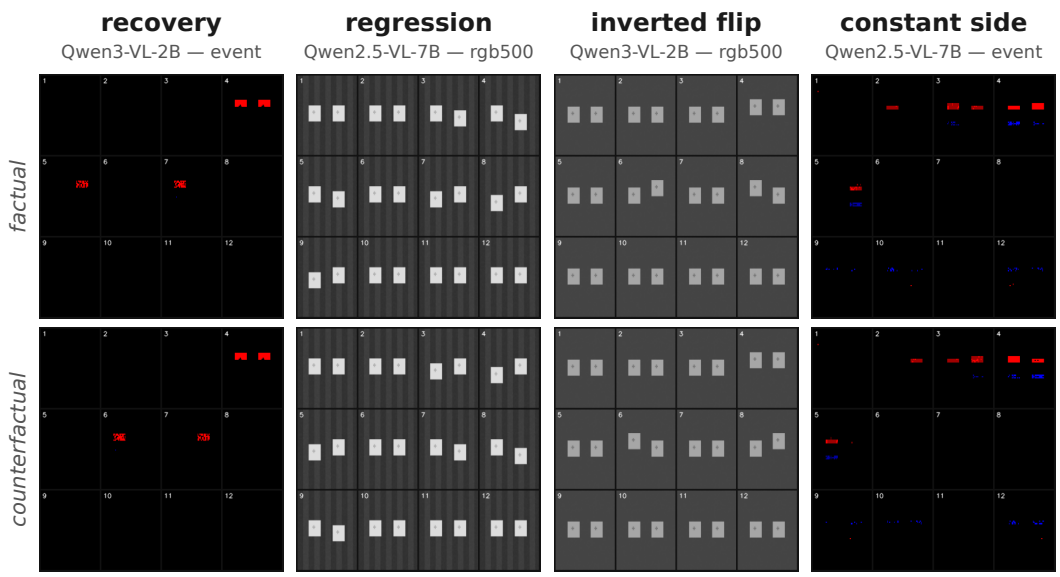


Figure 4: **Four pair-level error cases.** Each column shows a selected case, with the factual input above its counterfactual. Event contact sheets use red and blue for positive and negative events and white numbers for temporal order; RGB500 inputs show chronological frames. Predictions and gold labels appear in Table 5.

6 Failure Analysis and Reporting Boundaries

Pair-level failure modes. Near-50% sample accuracy has more than one interpretation: a constant-side predictor ignores the intervention and gets one answer per pair right, while an inverted flip changes the prediction between the two members in the direction opposite to the gold labels. We classify saved predictions as correct flip, inverted flip, constant left, constant right, or parse failure. Four cases selected by a deterministic rule appear in Table 5: recovery, regression, inverted flip, and persistent constant-side prediction. Figure 4 places their factual and counterfactual inputs side by side. The event slices and ordered 500-FPS frames show the available temporal evidence; the paired predictions show whether the model used it correctly.

Type	Model	Input	Pair	Gold	0-shot	Few-shot
recovery	Q3–2B	event	250	R/L	L/L	2-shot: R/L
regression	Q2.5–7B	rgb500	615	R/L	R/L	2-shot: L/L
inverted flip	Q3–2B	rgb500	250	R/L	L/L	2-shot: L/R
constant side	Q2.5–7B	event	0	L/R	L/L	2-shot: L/L

Table 5: **Representative pair-level behaviors.** Predictions are listed in factual/counterfactual order; the two gold answers are always opposites; L/R abbreviate left/right. The machine-readable supplement records sample identifiers, tasks, gold answers, raw outputs and selection rules.

Four saved examples. In the event-input recovery case, Qwen3-VL-2B answers left for both members at zero shots, against gold answers of right and left. Two demonstrations bring both predictions into line with the labels. The RGB500 regression goes the other way: Qwen2.5-VL-7B gets both answers right without demonstrations, then answers left twice with two demonstrations. Qwen3-VL-2B on RGB500 provides the inverted flip. Its two-shot predictions are left and right, the reverse of the right/left gold order. In the remaining event-input case, Qwen2.5-VL-7B stays with left/left at zero and two shots despite opposite gold labels.

Only the inverted-flip case contributes to flip consistency, although both failure types have zero pair accuracy. Recovery and regression illustrate opposite effects of the same demonstration count. These selected examples explain failure modes; their prevalence must be read from the complete-pair metrics.

Supplementary model coverage. The main few-shot table uses revision-pinned open checkpoints and one selection protocol. Supplementary evaluations broaden the model coverage, although their reported model names do not establish immutable upstream revisions or direct-provider identity. Both Claude Opus variants complete eight cells within the requested output budget: at eight shots, Claude-Opus-4.6 reaches 13.8% event pair accuracy and Claude-Opus-4.8 reaches 17.7% on events and 19.1% on RGB500. The 16 GPT cells in supplementary set B contain 2,000 samples each but violate output-budget or reasoning controls. They remain diagnostics in Appendix A. Across supplementary sets A and B, 112 of 139 formal-size metric files meet the registered output controls.

Completeness and comparability are separate requirements. A full prediction set enables pair scoring but does not repair a decoding-budget violation. Diagnostic results retain that designation even when they exceed compliant scores. We keep overlapping model names in supplementary sets A and B separate because their underlying identities are unverified.

Limits of cross-block comparison. The evaluated groups differ in representation and training history, precluding a controlled cross-family ranking. The frame-difference result shows that high-frame-rate RGB carries the answer when temporal change is exposed explicitly. Pair accuracy measures whether an interface yields correct responses; it does not identify why another interface fails.

7 Limitations and External Validity

Task scope. Scenes are procedurally generated at 64×64 pixels, with binary questions and minimal language. The benchmark measures responses to a controlled temporal change in this setting rather than open-ended question answering. One-sided temporal inputs attain 82.0–86.8% pair accuracy within policy, reinforcing our reading of high pair accuracy as intervention-consistent classification under the stated construction, not as evidence of compositional reasoning about events.

Simulation. Frozen v3 applies thinning after event simulation; without this matching, only 27.7% of v2e pairs have equal 32 ms signatures. The alternative-retention controls and independent-generator tests address dependence on that matching rule and on v2e, but cannot eliminate every shortcut in a procedurally generated scene. The OOD splits vary appearance or simulated sensor settings, and the second generator changes the synthetic event mechanism; none is a physical-sensor test. DSEC (Gehrig et al., 2021) has no compatible pair labels and serves only for unlabeled representation stress testing.

Interfaces and provenance. Event contact sheets and native video differ in presentation, so their scores do not isolate sensing modality; explicit RGB frame differences show that preprocessing also controls access to the signal. Although each media item carries 12 temporal units, prompts grow as demonstrations are added, so shot-count effects can include context-length effects. The open/local matrix and supplementary evaluations do not cover unreleased event-language checkpoints or verified direct-provider frontier evaluations. Supplementary model names are reported identities, with unverified upstream revisions. GPT evaluations that fail the requested output budget remain diagnostics.

8 Conclusion

The retention-policy controls, shortcut checks, and independent generator address RQ1 within the controlled synthetic setting: temporal information remains predictive across the tested representations and generators. Complete directional policy-transfer estimates are still unavailable. For RQ2, resolution and representation determine much of the gap between near-ceiling supervised models and the more variable VLM results; interface choice and demonstration count matter within the latter group. The answer to RQ3 is narrower: performance remains high under the tested scene shifts but varies more under sensor shifts, and none of these experiments establishes physical-sensor transfer.

EventTimeQA makes responses to temporal interventions measurable through paired counterfactuals. Its construction records, shortcut checks, and statistical procedures support interpretation of these paired results. Coarse matching erases the answer at 32 ms, whereas supervised encoders recover it almost perfectly at 2–4 ms. Reporting pair accuracy together with the representation is therefore essential; physical-sensor transfer and human performance remain open questions.

AI USE STATEMENT

Generative AI tools assisted with drafting the appendix clarifications, inspecting implementation details, and checking arithmetic consistency in the reported tables. No new model experiments were conducted for these appendix additions. The authors are responsible for verifying the resulting text and numerical corrections before submission.

REFERENCES

Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, Yao Dou, Jaden Park, Jianfeng Gao, Yong Jae Lee, and Jianwei Yang. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*, 2024.

Marco Cannici, Marco Ciccone, Andrea Romanoni, and Matteo Matteucci. A differentiable recurrent surface for asynchronous event-based data. In *European Conference on Computer Vision*, pp. 136–152. Springer, 2020.

Rohan Choudhury, Jean-Sebastien Dandurand, Kai Qiu, Kshitij Madhav Bhat, Kartik Sharma, Liza Dahiya, Yizhou Zhao, Souraja Kundu, Chun-Hsien Lin, Kris M Kitani, and László A. Jeni. Fps-bench: A benchmark for high frame-rate video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18598–18608, 2026.

Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44 (1):154–180, 2022.

Mathias Gehrig and Davide Scaramuzza. Recurrent vision transformers for object detection with event cameras. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13884–13893. IEEE, 2023.

Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, 2021.

Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic dvs events. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1312–1321. IEEE, 2021.

Pengteng Li, Yunfan Lu, Pinghao Song, Wuyang Li, Huizai Yao, and Hui Xiong. Eventvl: Understand event streams via multimodal large language model. *arXiv preprint arXiv:2501.13707*, 2025.

Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128$\times$128 120 dB 15μs latency asynchronous temporal contrast vision sensor. *IEEE journal of solid-state circuits*, 43(2):566–576, 2008.

Shaoyu Liu, Jianing Li, Guanghui Zhao, Yunjian Zhang, and Xiangyang Ji. Eventbench: Towards comprehensive benchmarking of event-based mllms. *arXiv preprint arXiv:2511.18448*, 2025a.

Shaoyu Liu, Jianing Li, Guanghui Zhao, Yunjian Zhang, Xin Meng, Fei Richard Yu, Xiangyang Ji, and Ming Li. Eventgpt: Event stream understanding with multimodal large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29139–29149. IEEE, 2025b.

Shaoyu Liu, Jianing Li, Guanghui Zhao, Yunjian Zhang, Wen Jiang, Ming Li, and Xiangyang Ji. EventFlash: Towards efficient MLLMs for event-based vision. *arXiv preprint arXiv:2602.03230*, 2026.

Dongyue Lu, Rong Li, Ao Liang, Lingdong Kong, Wei Yin, Lai Xing Ng, Benoit R Cottereau, Camille Simon Chane, and Wei Tsang Ooi. Eventdrive: Event cameras for vision-language driving intelligence. *arXiv preprint arXiv:2606.18242*, 2026.

Mohammad Mohammadi, Ziyi Wu, and Igor Gilitschenski. Tespec: Temporally-enhanced self-supervised pretraining for event cameras. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7782–7793. IEEE, 2025.

Garrick Orchard, Ajinkya Jayawant, Gregory K Cohen, and Nitish Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in neuroscience*, 9:437, 2015.

Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. Esim: an open event camera simulator. In *Conference on robot learning*, pp. 969–982. PMLR, 2018.

Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3852–3861. IEEE, 2019.

Hongwei Ren, Fei Ma, Xiaopeng Lin, Yuetong Fang, Hongxiang Huang, Yue Zhou, Yulong Huang, Haotian Fu, Ziyi Yang, Youxin Jiang, Xiangqian Wu, and Bojun Cheng. Scalable event cloud network for event-based classification. *arXiv preprint arXiv:2412.20803*, 2024.

Hongwei Ren, Youxin Jiang, Qifei Gu, and Xiangqian Wu. Generating adversarial events: A motion-aware point cloud framework. *arXiv preprint arXiv:2602.08230*, 2026a.

Hongwei Ren, Youxin Jiang, Tuopusen Huang, and Xiangqian Wu. A systematic survey on event camera representation learning. *arXiv preprint arXiv:2606.23078v1*, 2026b.

Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, and Ryad Benosman. Hats: Histograms of averaged time surfaces for robust event-based object classification. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 1731–1740. IEEE, 2018.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019.

Jianrui Zhang, Mu Cai, and Yong Jae Lee. Vinoground: Scrutinizing lmms over dense temporal reasoning with short videos. *arXiv preprint arXiv:2410.02763*, 2024.

Mingfang Zhang, Jingjing Pan, Ashutosh Kumar, Rajat Saini, Mustafa Erdogan, Hsuan-Kung Yang, Caixin Kang, Yifei Huang, Yoichi Sato, and Quan Kong. Cast-bench: Benchmarking causal chain-grounded spatio-temporal reasoning for video question answering. *arXiv preprint arXiv:2605.23216*, 2026.

Jiazhou Zhou, Xu Zheng, Yuanhuiyi Lyu, and Lin Wang. Eventbind: Learning a unified representation to bind them all for event-based open-world understanding. In *European Conference on Computer Vision*, pp. 477–494. Springer, 2024.

A Strict Multimodal Few-Shot Protocol

Only training-split pairs can become demonstrations. After filtering for complete pairs and balancing task and answer side, the selector fixes a nested sequence of four pairs. A k-shot prompt takes the first $k/2$ of them; neither member appears without its counterpart. Before inference, we check for shared pair IDs, sample IDs, or media paths between the query and its demonstrations, and verify 12 temporal units per media item. Selection uses the task named in the visible question, never the query label, pair ID, pixels, or model output. Decoding is greedy, and unparsable outputs count as failures. Table 6 contains the full metric matrix.

B Extended Few-Shot Model Coverage

Table 4 covers the revision-pinned Qwen and Gemma checkpoints run with one protocol commit. Other open and local VLMs, evaluated with the same budget and nested demonstrations, appear in Table 7. Supplementary set A adds 99 cells from reported DeepSeek, GLM, GPT, Moonshot, and Qwen models. DeepSeek-v4-Pro and GLM-5.2 reach 13.1% and 13.2% RGB500 pair accuracy at

Model	Input	Shots	Sample acc.	Pair acc.	Pair 95% CI	Flip cons.
Qwen3-VL-2B	event	0	50.1	0.3	[0.0,0.7]	0.3
Qwen3-VL-2B	event	2	51.7	4.8	[3.5,6.2]	6.2
Qwen3-VL-2B	event	4	50.1	1.3	[0.6,2.0]	2.3
Qwen3-VL-2B	event	8	50.0	0.1	[0.0,0.3]	0.1
Qwen3-VL-2B	rgb500	0	50.6	1.3	[0.6,2.0]	1.4
Qwen3-VL-2B	rgb500	2	50.2	0.8	[0.3,1.4]	1.2
Qwen3-VL-2B	rgb500	4	50.0	0.2	[0.0,0.5]	0.3
Qwen3-VL-2B	rgb500	8	52.5	5.3	[3.9,6.7]	5.6
Qwen2.5-VL-7B	event	0	50.0	0.0	[0.0,0.0]	0.0
Qwen2.5-VL-7B	event	2	50.0	0.0	[0.0,0.0]	0.0
Qwen2.5-VL-7B	event	4	50.0	0.0	[0.0,0.0]	0.0
Qwen2.5-VL-7B	event	8	50.0	0.0	[0.0,0.0]	0.0
Qwen2.5-VL-7B	rgb500	0	50.0	0.1	[0.0,0.3]	0.1
Qwen2.5-VL-7B	rgb500	2	50.0	0.0	[0.0,0.0]	0.0
Qwen2.5-VL-7B	rgb500	4	50.0	0.0	[0.0,0.0]	0.0
Qwen2.5-VL-7B	rgb500	8	50.0	0.0	[0.0,0.0]	0.0
Gemma-3-12B-it	event	0	50.25	1.4	[0.7,2.2]	2.3
Gemma-3-12B-it	event	2	47.55	4.3	[3.1,5.6]	13.5
Gemma-3-12B-it	event	4	47.35	5.0	[3.7,6.4]	15.3
Gemma-3-12B-it	event	8	46.80	4.7	[3.4,6.0]	15.8
Gemma-3-12B-it	rgb500	0	49.05	2.7	[1.7,3.8]	7.3
Gemma-3-12B-it	rgb500	2	49.80	1.0	[0.4,1.7]	2.4
Gemma-3-12B-it	rgb500	4	50.20	0.8	[0.3,1.4]	1.2
Gemma-3-12B-it	rgb500	8	49.80	0.7	[0.2,1.3]	1.8

Table 6: **Complete multimodal few-shot metrics.** All values are percentages on the same 2,000 ID samples (1,000 complete pairs). Pair intervals are percentile intervals from 10,000 complete-pair bootstrap resamples. Flip consistency is the fraction of pairs whose two predictions differ, whether the direction is correct or inverted.

their respective peaks. GPT-5.6-Luna reaches 88.4% in the RGB500 8-shot cell, but that evaluation violates the requested output and reasoning controls. The higher partial GPT-5.6-Sol scores have the same problem and are not validated benchmark scores.

Moonshot-8K cannot accept the original nine-image 8-shot prompt: at approximately 9.7K tokens, it exceeds the 8,192-token limit. A packed variant combines the eight selected demonstrations into four images while keeping 12 temporal units per example. It reaches 16.8% pair accuracy on event input and 7.1% on RGB500. Packing changes the image format, however, so these cells cannot support a within-format shot comparison. Remaining GPT cells were unavailable at the evidence freeze.

Supplementary set B contributes 40 cells with formal sample counts (Table 10). Both Claude Opus variants complete eight cells within the requested output budget. Claude-Opus-4.6 moves from 5.1% event pair accuracy at zero shots to 13.8% at eight, while its RGB500 result peaks at 13.4% with two shots. At eight shots, Claude-Opus-4.8 reaches 17.7% on events and 19.1% on RGB500. Gemini-3.1-Flash-Lite also completes eight compliant cells; its maxima are 31.9% on events and 74.0% on RGB500, both at two shots.

The GPT cells each contain 2,000 samples, yet all 16 fail the output-budget or reasoning controls. Overlapping cells from supplementary sets A and B are not averaged, and reported model names do not establish that the underlying implementations are identical. Set B's zero-shot event/RGB500 values—84.4%/97.6% for GPT-5.6-Sol and 45.4%/92.8% for GPT-5.6-Terra—remain diagnostics. Across the two sets, 112 of 139 formal-size metric files meet the registered output controls. Direct-provider provenance is unverified throughout. Small preliminary checks without complete test matrices are excluded from model comparison.

C Statistical and Reporting Details

Every confidence interval resamples complete pair IDs. For few-shot effects, predictions on identical test pairs enter pair-cluster sign-flip permutation tests and paired bootstrap intervals. The 12 planned

Model	Input	0-shot	2-shot	4-shot	8-shot
Qwen2.5-VL-7B	event	0.0 [0.0,0.0]	0.0 [0.0,0.0]	0.0 [0.0,0.0]	0.0 [0.0,0.0]
Qwen2.5-VL-7B	rgb500	**0.1 [0.0,0.3]**	0.0 [0.0,0.0]	0.0 [0.0,0.0]	0.0 [0.0,0.0]
Qwen3-VL-2B	event	0.3 [0.0,0.7]	**4.8 [3.5,6.2]**	1.3 [0.6,2.0]	0.1 [0.0,0.3]
Qwen3-VL-2B	rgb500	1.3 [0.6,2.0]	0.8 [0.3,1.4]	0.2 [0.0,0.5]	**5.3 [3.9,6.7]**
Qwen3.5-27B	event	8.8 [7.1,10.6]	10.8 [8.9,12.7]	7.9 [6.3,9.6]	**12.5 [10.5,14.6]**
Qwen3.5-27B	rgb500	16.1 [13.9,18.4]	45.6 [42.5,48.7]	51.0 [47.9,54.1]	**56.7 [53.6,59.8]**
gemma-3-12b-it	event	1.4 [0.7,2.2]	4.3 [3.1,5.6]	**5.0 [3.7,6.4]**	4.7 [3.4,6.0]
gemma-3-12b-it	rgb500	**2.7 [1.7,3.8]**	1.0 [0.4,1.7]	0.8 [0.3,1.4]	0.7 [0.2,1.3]
MiniCPM-V-4_5	event	**5.4 [4.1,6.8]**	0.0 [0.0,0.0]	2.6 [1.7,3.6]	0.2 [0.0,0.5]
MiniCPM-V-4_5	rgb500	**37.0 [34.0,40.0]**	0.9 [0.4,1.5]	8.3 [6.6,10.1]	3.7 [2.6,4.9]
InternVL3_5-8B	event	4.2 [3.0,5.5]	7.6 [6.0,9.3]	**10.3 [8.4,12.2]**	6.1 [4.7,7.7]
InternVL3_5-8B	rgb500	**53.8 [50.7,56.9]**	11.3 [9.4,13.2]	23.2 [20.6,25.9]	6.6 [5.1,8.2]

Table 7: **Extended open and local multimodal few-shot matrix.** Entries are ID pair accuracy in percent with pair-cluster 95% confidence intervals over 1,000 complete test pairs. All models use the same 12-unit evidence budget and nested training demonstrations. Bold marks the best shot count within each model-interface row.

Model	Input	0-shot	2-shot	4-shot	8-shot
deepseek-v4-flash	event	0.0 [0.0,0.0]	3.3 [2.2,4.4]	**8.7 [7.0,10.5]**	0.2 [0.0,0.5]
deepseek-v4-flash	rgb500	1.4 [0.7,2.2]	**5.0 [3.7,6.4]**	4.3 [3.1,5.6]	1.8 [1.0,2.7]
deepseek-v4-pro	event	5.7 [4.3,7.2]	**6.9 [5.3,8.5]**	0.9 [0.4,1.5]	0.0 [0.0,0.0]
deepseek-v4-pro	rgb500	0.9 [0.4,1.5]	6.4 [4.9,7.9]	**13.1 [11.0,15.2]**	0.1 [0.0,0.3]
glm-5.2	event	5.8 [4.4,7.3]	**12.4 [10.4,14.4]**	5.6 [4.2,7.1]	0.1 [0.0,0.3]
glm-5.2	rgb500	1.1 [0.5,1.8]	9.7 [7.9,11.6]	**13.2 [11.1,15.3]**	0.9 [0.4,1.5]
GPT-5.6-Luna	event	7.3 [5.7,8.9]	2.8 [1.8,3.9]	**17.1 [14.8,19.4]**	2.2 [1.4,3.2]
GPT-5.6-Luna	rgb500	11.0 [9.1,13.0]	8.8 [7.1,10.6]	87.3 [85.2,89.3]	**88.4 [86.4,90.3]**
GPT-5.6-Sol	rgb500	**99.1 [98.5,99.6]**	94.9 [93.5,96.2]	—	—
GPT-5.6-Terra	event	—	—	—	**63.1 [60.1,66.0]**

Table 8: **Supplementary set A, part I (GLM, DeepSeek, GPT).** Entries are ID pair accuracy in percent with pair-cluster 95% confidence intervals over 1,000 complete test pairs. Reported model identities and upstream revisions are unverified. Dashes denote unavailable cells. GPT outputs violate the requested output or reasoning controls and remain diagnostic; their scores are not validated benchmark results. These evaluations provide supplementary coverage and do not underpin the controlled-synthetic main claims.

Model	Input	0-shot	2-shot	4-shot	8-shot
moonshot-v1-128k-vision-preview	event	0.0 [0.0,0.0]	8.2 [6.5,9.9]	6.9 [5.4,8.5]	**13.1 [11.1,15.2]**
moonshot-v1-128k-vision-preview	rgb500	0.0 [0.0,0.0]	**5.5 [4.1,7.0]**	0.9 [0.4,1.5]	5.3 [4.0,6.7]
moonshot-v1-32k-vision-preview	event	0.0 [0.0,0.0]	8.2 [6.6,9.9]	7.6 [6.0,9.3]	**12.6 [10.6,14.7]**
moonshot-v1-32k-vision-preview	rgb500	0.0 [0.0,0.0]	4.7 [3.4,6.1]	0.8 [0.3,1.4]	**6.7 [5.1,8.3]**
moonshot-v1-8k-vision-preview	event	0.0 [0.0,0.0]	8.0 [6.4,9.8]	7.0 [5.4,8.6]	**16.8 [14.5,19.1]**
moonshot-v1-8k-vision-preview	rgb500	0.0 [0.0,0.0]	5.0 [3.7,6.4]	0.8 [0.3,1.4]	**7.1 [5.5,8.7]**
qwen3-235b-a22b	event	0.0 [0.0,0.0]	0.8 [0.3,1.4]	0.8 [0.3,1.4]	0.0 [0.0,0.0]
qwen3-235b-a22b	rgb500	0.0 [0.0,0.0]	0.0 [0.0,0.0]	**3.2 [2.2,4.3]**	0.0 [0.0,0.0]
qwen3.6-27b	event	9.4 [7.6,11.3]	11.4 [9.4,13.4]	10.0 [8.2,11.9]	**15.6 [13.4,17.9]**
qwen3.6-27b	rgb500	18.5 [16.1,21.0]	35.6 [32.7,38.5]	30.4 [27.6,33.2]	**40.8 [37.8,43.8]**
qwen3.6-35b-a3b	event	5.5 [4.1,7.0]	10.2 [8.4,12.1]	7.4 [5.8,9.1]	**12.5 [10.5,14.6]**
qwen3.6-35b-a3b	rgb500	10.2 [8.4,12.1]	37.1 [34.1,40.1]	**40.6 [37.6,43.6]**	32.9 [30.0,35.9]
qwen3.6-plus	event	4.8 [3.5,6.2]	16.7 [14.4,19.0]	21.1 [18.6,23.6]	**24.3 [21.6,27.1]**
qwen3.6-plus	rgb500	27.8 [25.0,30.6]	**45.6 [42.5,48.7]**	36.2 [33.3,39.2]	38.3 [35.2,41.4]
qwen3.7-plus	event	20.6 [18.2,23.2]	23.9 [21.2,26.5]	22.7 [20.1,25.4]	**28.9 [26.1,31.8]**
qwen3.7-plus	rgb500	29.0 [26.2,31.9]	**48.9 [45.9,52.0]**	36.3 [33.3,39.3]	38.1 [35.1,41.1]

Table 9: **Supplementary set A, part II (Moonshot, Qwen).** Entries are ID pair accuracy in percent with pair-cluster 95% confidence intervals over 1,000 complete test pairs. Reported model identities and upstream revisions are unverified. Moonshot 8K's nine-image 8-shot prompt exceeds its context limit, so eight demonstrations are packed into four worked-example images plus the query. Each example retains 12 temporal units, and selection and output controls are unchanged. These packed values do not support a within-format shot comparison.

Qwen contrasts are corrected within model–interface families with Holm family-wise and Benjamini–Hochberg false-discovery procedures. Gemma-3-12B-it contrasts form separate families. Machine-readable reports retain estimates, effect sizes, raw and adjusted p values, seeds, and the count of 10,000 resamples.

Model	Input	0-shot	2-shot	4-shot	8-shot
Claude-Opus-4.6	event	5.1 [3.8,6.5]	12.1 [10.1,14.1]	12.5 [10.5,14.6]	**13.8 [11.7,16.0]**
Claude-Opus-4.6	rgb500	8.9 [7.2,10.7]	**13.4 [11.3,15.5]**	13.0 [10.9,15.1]	10.6 [8.7,12.6]
Claude-Opus-4.8	event	14.2 [12.1,16.4]	14.1 [12.0,16.3]	11.5 [9.5,13.6]	**17.7 [15.4,20.1]**
Claude-Opus-4.8	rgb500	13.4 [11.3,15.6]	11.4 [9.5,13.4]	14.5 [12.4,16.7]	**19.1 [16.7,21.6]**
Gemini-3.1-Flash-Lite	event	25.9 [23.2,28.6]	**31.9 [29.0,34.8]**	31.8 [29.0,34.7]	30.4 [27.5,33.3]
Gemini-3.1-Flash-Lite	rgb500	68.0 [65.0,70.9]	**74.0 [71.2,76.7]**	68.0 [65.1,70.9]	72.7 [69.9,75.5]
GPT-5.6-Sol	event	**84.4 [82.1,86.6]**	64.7 [61.7,67.7]	62.4 [59.4,65.5]	58.9 [55.9,61.9]
GPT-5.6-Sol	rgb500	**97.6 [96.6,98.5]**	86.6 [84.4,88.7]	89.3 [87.4,91.2]	92.8 [91.2,94.4]
GPT-5.6-Terra	event	**45.4 [42.3,48.5]**	30.4 [27.6,33.3]	32.3 [29.4,35.2]	30.7 [27.9,33.6]
GPT-5.6-Terra	rgb500	**92.8 [91.2,94.4]**	80.1 [77.6,82.6]	86.3 [84.1,88.5]	89.1 [87.1,91.0]

Table 10: **Supplementary set B (Claude, Gemini, GPT).** Entries are ID pair accuracy in percent with pair-cluster 95% confidence intervals over 1,000 complete test pairs. Claude and Gemini conditions satisfy the requested eight-token, no-reasoning budget. Every GPT condition violates the output or reasoning controls and is diagnostic only. Reported model names do not verify upstream identities or immutable revisions. Overlapping names in sets A and B do not justify merging their measurements.

D Evidence Scope

Supplementary results are grouped by evaluation set, and their reported model names are descriptive labels rather than independently verified upstream identities. A complete sample count does not imply protocol compliance. Tables 8, 9, and 10 distinguish the compliant conditions from diagnostic GPT outputs. Physical-sensor capture and human evaluation are outside the completed evidence. No result in these tables establishes transfer to real sensors or human-level temporal perception.

E Retention-Policy Audit

Policy	Exact before	Retained	Metadata	Count	No time	Temporal pair
Unmodified	27.7	100.00	50.00	55.99	56.89	92.81
First retained	27.7	94.72	50.00	50.00	50.00	97.01
Random retained	27.7	94.72	50.00	50.00	50.00	95.81

Table 11: **Retention-policy audit on 1,000 pairs per policy.** Values are percentages. "Exact before" measures equality of unmodified 32 ms signatures, and "Retained" is the mean fraction of events kept; across the retained streams the per-pair minimum is 56.32%. The metadata, count, and no-time columns give ID sample accuracy for nuisance-only ridge controls. "Temporal pair" gives within-policy pair accuracy for the bilateral temporal-histogram control; its 95% pair-cluster intervals are $[88.62, 96.41]$, $[94.01, 99.40]$, and $[92.22, 98.80]$ for the unmodified, first-retained, and random-retained streams.

Table 11 reports the audit behind §3 on 1,000 pairs per policy: equality of unmodified 32 ms signatures before thinning, the fraction of events retained, the nuisance-only ridge controls, and the bilateral temporal positive control. These are the values quoted in §3 for the 27.7% signature-equality rate, the 94.72% mean retention with a 56.32% per-pair minimum, and the 55.99% and 56.89% count-only and no-time controls on unmodified events.

The 95% interval $[88.62, 96.41]$ belongs to the unmodified within-policy temporal control. It is an uncertainty interval around one estimate, not a range of observed transfer accuracies. Per-direction train-policy $\rightarrow$ test-policy cells were not individually measured in the available summary; no complete directional transfer matrix can be inferred from this table.