Research Report

1. Executive Overview

The challenge of evaluating whether multimodal large language models (MLLMs) genuinely reason over video content—rather than exploiting superficial shortcuts—has emerged as a central concern in the video understanding field. Current benchmarks report impressive accuracy numbers, yet careful analysis reveals that these scores frequently reflect exploitation of language priors, single-frame cues, and local visual patterns rather than authentic temporal comprehension. A presentation on benchmark design for long-video reasoning illuminated three critical question families that expose genuine capability gaps: event linking (connecting moments belonging to the same underlying process), state tracking (maintaining what changed, remained available, and is pending), and counterfactual/contrastive reasoning (understanding why this answer rather than a plausible alternative). The literature analysis confirms that the gap between reported benchmark performance and genuine video understanding capability is substantial—Gemini-3-Pro achieves 66.1% per-question accuracy on Video-MME-v2 but drops to 49.4% under group-based nonlinear scoring that penalizes inconsistent responses across related questions.

The most pressing unresolved bottleneck is the absence of evaluation methodologies that can distinguish genuine temporal reasoning from shortcut exploitation. VCBench demonstrates that even frontier models fail at basic counting tasks when queried at streaming timepoints, with prediction trajectories showing monotonicity violations and delayed responses to state changes. ReXTime reveals the largest human-model gaps (14.3%) for reasoning-across-time tasks where questions and answers occur in different video segments. Meanwhile, dedicated long-video models (LLaMA-VID, MovieChat, LWM) perform near-random on LVBench and Neptune, while simpler approaches with more frames outperform them—suggesting fundamental architectural limitations rather than mere processing constraints. The meeting discussion and literature synthesis together point toward a research agenda that prioritizes rigorous evaluation methodology, architectural innovation for genuine long-context reasoning, synthetic data approaches with guaranteed quality, and budget-aware deployment strategies for practical applications.

2. Problem Setting and Source Context

The source material is a presentation on designing benchmarks for evaluating whether models genuinely reason over long video content. The core problem addressed is that current video benchmarks can be partially solved via shortcuts: language priors allow models to answer questions without watching video, single key frames capture the essential information, and local cues near clip endings provide sufficient evidence. These shortcuts enable models to appear capable without genuine temporal understanding. The work focuses on tasks where

evidence is distributed over time and cannot be answered from a single salient frame.

Three question families structure the evaluation approach. Event linking connects moments belonging to the same underlying process, requiring models to track causal or sequential relationships across video segments. State tracking maintains what changed, what remained available, and what is pending throughout video playback. Counterfactual/contrastive reasoning requires understanding why this answer rather than a plausible alternative. These families represent increasingly challenging levels of temporal comprehension.

Annotation challenges are substantial. Moving from local recognition to temporal structure makes annotation difficult. Defining what counts as step start, failed attempts, or incidental motion requires careful protocol design. Annotators must separate temporal anchoring from semantic labeling. The meeting discussion emphasizes that hybrid evaluation design—combining short rationale/evidence trace generation with constrained answer format—exposes whether the model looked in the right place for the right reason. Rationale is scored against evidence anchors, not prose quality.

Two context budget settings are proposed. Fixed-budget evaluation applies the same visual budget to all models, enabling fair comparison. Adaptive evaluation allows models to retrieve more evidence but penalizes resource consumption, better reflecting deployment scenarios where context budget changes the nature of the task. Key findings include that current systems excel at identifying salient events but struggle with maintaining coherent state over distributed temporal evidence, and that stronger language models improve answer plausibility while potentially hiding grounding failures. Human performance remains robust under imperfect sampling when temporal anchors are available, but drops sharply when sampling becomes too sparse.

3. Grounded Findings from the Source Material

3.1 Shortcut Exploitation in Current Benchmarks

The presentation identified systematic shortcut vulnerabilities in existing video understanding benchmarks. Language priors allow models to answer questions correctly without processing visual content at all—Llama 3-70B achieves 38% on video QA versus 25% random baseline, confirming significant textual bias. Single-frame baselines on LVBench show minimal performance degradation compared to full-video processing, indicating that temporal information is often not required for correct answers. Single-frame and video-only baselines frequently achieve optimal performance on MVBench, demonstrating that impressive benchmark numbers often reflect priors rather than genuine temporal understanding.

3.2 Three Question Families for Temporal Reasoning

Event linking requires connecting moments belonging to the same underlying process, testing whether models can identify that separate video segments

describe parts of a continuous causal or sequential chain. State tracking requires maintaining world state throughout video playback—tracking what changed, what remained available, and what is pending. VCBench's streaming multi-point query design reveals that prediction trajectories show monotonicity violations and delayed responses to state changes, indicating fundamental limitations in current architectures. Counterfactual/contrastive reasoning requires understanding why this answer rather than a plausible alternative, testing whether models can distinguish between superficially similar scenarios that differ in critical temporal details.

3.3 Hybrid Evaluation Design

The proposed evaluation combines short rationale/evidence trace generation with constrained answer format. The rationale exposes whether the model looked in the right place for the right reason. Critical insight: rationale is scored against evidence anchors, not prose quality. This approach separates genuine reasoning from confident confabulation—stronger LLMs can make outputs sound reasonable while temporal evidence remains wrong, so evaluation must verify the evidence rather than trusting the conclusion.

3.4 Context Budget as Task Variable

Two settings are proposed for context budget evaluation. Fixed-budget evaluation applies the same visual budget to all models, enabling fair capability comparison under resource constraints. Adaptive evaluation allows models to retrieve more evidence but applies metric penalties for additional resource consumption. The adaptive setting better reflects deployment scenarios where context budget is a real operational constraint rather than a nuisance variable. The meeting discussion emphasized that context budget is not just a nuisance variable—it changes the nature of the task, potentially making some reasoning approaches infeasible that would work with unlimited context.

3.5 State Maintenance Limitations

Current systems excel at identifying salient events but struggle with maintaining coherent state over distributed temporal evidence. VCBench demonstrates that even state-of-the-art models fail to maintain consistent world state across video playback, particularly for periodic event counting. The streaming query design exposes trajectory inconsistencies invisible to single-query benchmarks. This represents a critical capability gap for applications requiring tracking object identities, event counts, or cumulative information across video duration.

3.6 Protocol Design and Annotation Challenges

The meeting emphasized that protocol design matters as much as annotation volume. Defining what counts as step start, failed attempts, or incidental motion requires careful protocol design. Annotators must separate temporal anchoring from semantic labeling. Preserving ambiguity in metadata rather than forcing fake consensus may be more valuable than forcing annotation consensus. The annotation process must distinguish between recognizing an event occurred and correctly reasoning about its temporal relationships to other events.

3.7 Benchmark Lifespan Concerns

Popular benchmarks get optimized and stop being informative. Model progress becomes conflated with annotation design progress. This creates a dynamic where reported improvements may partially reflect benchmark-specific optimizations rather than genuine capability advances. The meeting suggested designing benchmarks as evolving frameworks rather than frozen leaderboards to maintain their diagnostic value over time.

4. Literature-Based Deep Analysis

4.1 Preserved Detailed Paper Analyses

VCBench: Streaming Counting for Spatial-Temporal State Maintenance

Problem and Task Setting: VCBench addresses the gap in evaluating how models maintain world state throughout video playback. Unlike existing benchmarks that present single queries at video end, VCBench uses streaming multi-point queries to observe prediction trajectories over time. The benchmark decomposes spatial-temporal state maintenance into 8 fine-grained subcategories: object counting (O1-Snap, O1-Delta, O2-Unique, O2-Gain) and event counting (E1-Action, E1-State, E2-Periodic, E2-Span).

Methodology: The benchmark employs three complementary evaluation metrics: GPA (Gaussian Precision Accuracy) for numerical precision, MoC (Monotonicity Consistency) for trajectory logical self-consistency, and UDA (Update Direction Accuracy) for temporal awareness. For O2-Periodic, loop concatenation generates 2-3 minute videos from TOMATO periodic actions. Offline evaluation uses OVO-Bench protocol: video truncated to query moment for each timepoint.

Main Evidence: VCBench contains 406 videos with frame-by-frame annotations of 10,071 event moments and 4,576 streaming query points across 1,000 QA pairs. Evaluation reveals significant deficiencies in state maintenance, particularly for periodic event counting. Online models (StreamingVLM) are currently under-explored, with only one model evaluated.

Relevance to Topic: VCBench directly addresses temporal state tracking, revealing that even state-of-the-art models fail to maintain consistent world state across video playback. The streaming query design exposes trajectory inconsistencies invisible to single-query benchmarks.

Limits/Caveats: Offline evaluation approximates streaming processing through video truncation, potentially overestimating performance since models can re-examine historical content. The dataset scale is limited (406 videos) due to high annotation costs for frame-by-frame temporal labeling.

ReXTime: Reasoning-Across-Time Benchmark

Problem and Task Setting: ReXTime focuses on the under-explored capability of temporal reasoning when questions and answers occur in different video

segments. The benchmark evaluates both VQA accuracy and moment localization, covering four event relation types: sequential, cause-effect, means-to-end, and parallel relationships.

Methodology: An LLM-assisted pipeline reduces annotation costs from $300 to $135 per 1,000 QA pairs. Human annotators curate all samples for accuracy and relevance. Fine-tuning experiments use LoRA adaptation on VTimeLLM stage-3 weights with the generated training dataset (9,695 samples).

Main Evidence: Human performance reaches 88.0% VQA accuracy versus GPT-4o's 73.7% (14.3% gap). Moment localization shows even larger gaps: Human 62.85% vs GPT-4o 34.00% at IoU≥0.5. Claude3-Opus achieves only 13.67% moment localization accuracy.

Relevance to Topic: Cross-segment temporal reasoning represents the frontier of video understanding capability, with cause-effect relationships particularly challenging. The human-model gap suggests significant room for architectural innovation.

Limits/Caveats: Fine-tuning gains are demonstrated but may not transfer to diverse video domains. The benchmark uses ActivityNet videos, which may have different characteristics from other video sources.

MVP: Shortcut-Aware Video-QA Benchmark

Problem and Task Setting: MVP addresses benchmark integrity by requiring models to correctly answer both videos in a minimal-change pair—visually similar videos with identical questions but opposing answers. This penalizes models exploiting shortcuts based on superficial visual or textual cues.

Methodology: The benchmark collects 55K multiple-choice video QA examples from 9 data sources spanning egocentric/exocentric videos, robotic interaction, and intuitive physics. Minimal-change pair scoring requires both pair members correct for positive credit. Data curation uses automated pipelines with human verification.

Main Evidence: Language-only baseline (Llama 3-70B) achieves 38% vs 25% random (+8pp language prior). Video-only baseline achieves 50% on random pairing but only 27.3% on minimal-change pairs. Best open-source model (LLaVA-OneVision) achieves 40.2% vs 92.9% human. Table 1 shows optimal MVBench performance frequently achieved through single-frame, video-only, or text-only shortcuts.

Relevance to Topic: MVP exposes systematic shortcut vulnerabilities that inflate benchmark scores, demonstrating that impressive-looking numbers often reflect priors rather than genuine temporal understanding.

Limits/Caveats: The benchmark may be too strict for some questions where multiple valid answers exist. CoT reasoning was disabled during evaluation, which may limit model capability.

Video-MME-v2: Comprehensive Video Understanding with Group-Based Evaluation

Problem and Task Setting: Video-MME-v2 advances evaluation methodology through a tri-level capability hierarchy (Information Aggregation, Temporal Dynamics, Complex Reasoning) and group-based nonlinear scoring that penalizes fragmented or guess-based correctness. The benchmark evaluates 800 videos with 3,200 questions.

Methodology: Group-based evaluation assesses both consistency (breadth across related questions) and coherence (depth in multi-step reasoning). Non-linear scoring requires all correlated questions correct within a group for full credit, assigning zero when any question in a consistency group is answered incorrectly. The evaluation involves 12 annotators and 50 reviewers with 3,300 human-hours.

Main Evidence: Human expert baseline: 90.7 Non-Lin Score, 94.9% Avg Acc. Gemini-3-Pro: 49.4 Non-Lin Score vs 66.1% Avg Acc (25% gap from nonlinear scoring). Robustness ratio (Non-Lin/Avg) reveals Gemini-3-Pro at 75% but smaller models at ~40%, indicating weaker consistency. Thinking-based reasoning improves with subtitles (+11.2 for Gemini-3-Pro) but degrades without textual cues.

Relevance to Topic: This benchmark demonstrates that evaluation methodology critically affects conclusions—standard per-question accuracy significantly overestimates actual reliability and consistency in video understanding.

Limits/Caveats: The benchmark focuses on publicly available videos which may have different characteristics from private or specialized content. Human evaluation may have different failure modes than model evaluation.

LVBench: Extreme Long Video Understanding

Problem and Task Setting: LVBench specifically targets videos spanning several hours (up to several hours), addressing embodied intelligence, movie analysis, and live sports scenarios where short-video benchmarks fail. The benchmark defines six core capabilities: entity recognition, key information retrieval, temporal grounding, summarization, comparison, and correlation.

Methodology: Videos are collected from public sources with annotations through manual effort and model assistance. An LLM filtering step removes questions answerable from text alone—after filtering, LLaVA-NeXT dropped from 48.9% to 32.2% (16.7pp reduction). Evaluation uses 1 FPS frame extraction for long video processing.

Main Evidence: Human performance: 94.4% average accuracy. Gemini 1.5 Pro: 33.1% overall. LLaMA-VID, MovieChat, LWM nearly equivalent to random selection. LLaVA-NeXT with 32 input frames outperformed all long-video native models except Gemini 1.5 Pro. Gemini 1.5 Pro achieved highest accuracy on sports (41.2%) but lowest on documentaries (25.4%).

Relevance to Topic: LVBench quantifies severe degradation in video understanding as duration increases, revealing that architectural solutions for long context have not translated to practical capability improvements.

Limits/Caveats: Audio data is excluded despite providing valuable context. Most models lack effective audio processing capabilities. Long-video native models may have different failure modes not captured by current evaluation.

Neptune: Scalable Long Video Benchmark

Problem and Task Setting: Neptune addresses scalability in long video benchmark creation through a semi-automatic pipeline leveraging VLMs and LLMs for annotation, with human verification as quality control. The benchmark includes Neptune-Full (3,268 QADs across 2,405 videos) and Neptune-MMH (1,171 QADs emphasizing visual reasoning).

Methodology: Pipeline stages: video selection from YT-Temporal-1Bn, signal extraction (PaLI-3 captions, ASR, metadata, shot boundaries), automatic caption generation, QAD generation via prompted LLMs, and human verification. The scalable approach halves annotation effort compared to fully manual methods. GEM (Gemma Equivalence Metric) provides open-ended evaluation scoring.

Main Evidence: Gemini-1.5-Pro with all frames + ASR: 45.05% on Neptune-Full, 31.9% on Neptune-MMH. Single middle frame (55.57%) outperforms first frame (42.26%). Neptune-MMH (visual-focused) shows larger model advantage than Neptune-Full (multimodal). Most benchmarks saturate at ~50 frames; Neptune requires >150 frames for continued improvement, indicating genuine long-context challenge.

Relevance to Topic: Neptune reveals that ASR-heavy content dominates many "long video" datasets, with visual understanding contributing minimally. True long-video benchmarks must emphasize vision-dependent questions to avoid measuring ASR capability rather than visual understanding.

Limits/Caveats: The dataset may inherit biases from the Gemini model used for generation. Neptune-MMH still contains some audio-dependent questions despite filtering.

MMOU: Omni-Modal Long Video Understanding

Problem and Task Setting: MMOU evaluates joint audio-visual reasoning across 15,000 questions paired with 9,038 videos spanning 13 skill categories requiring tightly coupled multimodal understanding. The benchmark requires

both modalities to answer questions correctly, not just single-modality sufficiency.

Methodology: All questions are manually annotated across multiple turns by professional annotators. Skill categories include temporal understanding, counting, needle-in-the-haystack reasoning, and inference. Evaluation includes both closed-source (Gemini 2.5 Pro) and open-source models (Qwen3-Omni, MiniCPM-o).

Main Evidence: Gemini 2.5 Pro: 64.2% overall (84.3% human). Qwen3-VL-32B video-only: 44% (confirms cross-modal necessity). Audio-only models show 17.7-35.6% performance drop. Temporal position sensitivity: sharp accuracy drop when evidence appears later in video. Cross-modal models (Qwen3-Omni-30B: 46.0%) significantly outperform video-only models (Qwen3-VL-32B: 44.0%).

Relevance to Topic: MMOU demonstrates that audio-visual integration is critical for realistic video understanding, and that performance degrades significantly when relevant information appears late in long videos.

Limits/Caveats: The benchmark uses publicly available web videos which may have content biases. Multiple-choice evaluation may not capture full open-ended reasoning capabilities.

BenchScope: Benchmark Redundancy Analysis

Problem and Task Setting: BenchScope introduces Effective Dimensionality (ED) to measure independent signal diversity in evaluation suites, addressing the question of whether benchmarks actually provide independent information beyond what existing evaluations measure.

Methodology: ED computes the participation ratio of a centered benchmark-score spectrum. Higher ED indicates more independent measurement axes. The analysis uses per-instance granularity, revealing 5× more redundancy than per-category analysis. ED-Greedy algorithm selects tasks to maximize measurement diversity.

Main Evidence: Open LLM Leaderboard (6 scores): ED=1.7 effective dimensions. BBH-MMLU-Pro correlation: $\rho=0.96$ (near-identical signals). Measurement breadth varies 20× across benchmarks. Random Dirichlet weighting changes champion 38% of the time. Weight sensitivity: even moderate perturbations ($\alpha=5$) change champion 19% of the time.

Relevance to Topic: BenchScope provides methodological tools for evaluating whether new benchmarks actually add independent signals, essential for understanding what new video benchmarks contribute beyond existing evaluations.

Limits/Caveats: ED is population-conditional—restricting model set changes ED values. Binary spectra systematically overestimate absolute dimensionality. Negative correlations are specific to curated 188-model subset.

Triage: Hierarchical Visual Budgeting

Problem and Task Setting: Triage addresses video processing efficiency through hierarchical resource allocation—first selecting keyframes (Frame-Level Budgeting), then allocating tokens within selected frames (Token-Level Budgeting)—to reduce computational requirements while maintaining or improving performance.

Methodology: Frame importance scoring combines scene change (adjacent frame similarity), motion intensity (pixel difference), and text relevance (CLIP similarity). Adaptive temporal bucketing ensures narrative preservation by distributing selection across video timeline. Token-level uses Core Tokens (high relevance) plus Context Tokens via MMR (Maximal Marginal Relevance) for diversity.

Main Evidence: 50% token retention: Triage 60.4 vs PyramidDrop 58.6, FastV 59.0, DyCoke 59.1 on Video-MME. 25% Triage outperforms 50% for competing methods on LongVideoBench. LVBench shows highest gains from frame-level prioritization. Ablation confirms both stages necessary: Frame w/o Token 58.3%, Token w/o Frame 59.4%, Full 60.1%.

Relevance to Topic: Triage demonstrates that intelligent visual budgeting—selecting what to process rather than processing everything—improves both efficiency and accuracy by filtering noise and focusing computational resources.

Limits/Caveats: Performance gains may vary for different video genres or query types. The hyperparameter sensitivity analysis suggests careful tuning may be required for optimal results.

ContextBudget: Budget-Aware Context Management

Problem and Task Setting: ContextBudget applies reinforcement learning to optimize context compression decisions in long-horizon agent reasoning, with progressive budget curriculum training that adapts to varying context constraints.

Methodology: BACM (Budget-Aware Context Management) formulates compression as a budget-constrained sequential decision problem. Budget-conditioned inference exposes remaining context headroom before appending observations. Commit-block aggregation enables adaptive compression timing and amount. BACM-RL uses GRPO with trajectory-level advantage broadcasting and curriculum stages (B*max to B*max/4).

Main Evidence: 32-objective regime: BACM-RL 4.545 vs MEM1 0.909 (5.0× improvement). 8k budget BACM-RL-30B (0.147) outperforms 128k Qwen3-235B-Inst (0.136) on BrowseComp-Plus. Performance remains stable across budget

reductions from 16k to 4k. RL ablation confirms importance: BACM w/o RL 0.208 vs BACM-RL 4.545 in 32-objective regime.

Relevance to Topic: ContextBudget shows that adaptive context management enables robust long-horizon reasoning under budget constraints, critical for deploying VLMs in resource-constrained video applications.

Limits/Caveats: RL training introduces complexity and may require careful hyperparameter tuning. The approach was evaluated on QA and browsing tasks; video-specific adaptation may be needed.

TimE: Multi-Level Temporal Reasoning Benchmark

Problem and Task Setting: TimE evaluates temporal reasoning across three hierarchical levels with 38,522 QA pairs: Level 1 (Basic Temporal Understanding: extraction, localization, computation, duration comparison, order comparison), Level 2 (Temporal Expression Reasoning: explicit, order, relative), Level 3 (Complex Temporal Relationship Reasoning: co-temporality, timeline, counterfactual).

Methodology: Three sub-datasets cover different challenges: TimE-Wiki (knowledge-intensive), TimE-News (rapid evolution), TimE-Dial (social interactions). QA synthesis combines rule-based templates with DeepSeek-V3/R1 models. RAG evaluation uses BM25, Vector, and Hybrid retrieval strategies. TimE-Lite (938 samples) provides human-annotated quality control.

Main Evidence: o3-mini: 52.62% on Order Reasoning, 54.34% on Co-temporality. Basic retrieval (Level 1): ~80% across 4 tasks. Complex reasoning (Level 3): 30-50% range. Test-time scaling improves reasoning but may hurt time retrieval. Deepseek-R1-Distill-Qwen-14B outperforms Qwen2.5-14B-Instruct on Order Compare and Duration Compare but underperforms on Extract and Localization.

Relevance to Topic: TimE quantifies the substantial gap between basic temporal retrieval and complex temporal reasoning, with LLMs performing well on extraction but struggling with ordering, causation, and counterfactual inference.

Limits/Caveats: The benchmark focuses on text-based temporal reasoning; video temporal reasoning may have different characteristics. RAG retrieval quality affects performance on TimE-News and TimE-Dial.

TEMPURA: Temporal Event Masked Prediction

Problem and Task Setting: TEMPURA addresses fine-grained temporal understanding by unifying masked event prediction (reconstructing missing events from context) with video segmentation (partitioning into temporally grounded event sequences). The approach trains on VER, a dataset of 500K videos with dense event captions.

Methodology: Two-stage training: Stage 1 uses masked event prediction (Fill-in-the-Middle paradigm) where model predicts pseudo-events and reasoning steps from surrounding context. Stage 2 focuses on video segmentation and dense captioning, partitioning videos into non-overlapping events with timestamps. VER construction uses GPT-4o for segmentation and LLM annotation.

Main Evidence: Charades-STA: TEMPURA 39.2 mIoU vs baseline 32.9 (+6.3). QVHighlights: TEMPURA 51.7 HIT@1 vs baseline 44.8 (+6.9). Ablation confirms both stages necessary: masked event prediction alone 35.8 mIoU, segmentation alone 36.1 mIoU, combined 39.2 mIoU.

Relevance to Topic: TEMPURA demonstrates that causal event understanding and fine-grained temporal segmentation are complementary capabilities that together improve temporal grounding beyond either alone.

Limits/Caveats: VER contains 500K training videos but evaluation uses standard benchmarks which may not fully cover the capabilities learned. The approach requires specific training; zero-shot transfer may be limited.

SynRL: Synthetic Temporal Primitives

Problem and Task Setting: SynRL addresses two critical limitations in video training data: (1) lack of temporal-centricity where answers can be inferred from isolated frames, and (2) systematic errors in proprietary model annotations. The approach teaches temporal primitives (direction, speed, state tracking) through programmatically generated synthetic videos with guaranteed ground-truth.

Methodology: Short-term videos (12 motion types): collision counting, direction identification, trajectory shape recognition, speed perception, motion counting, attribute change detection, rotation perception, relative position tracking, acceleration detection, velocity comparison, distance estimation, sequential event ordering. Long-term videos: state tracking with invisible intermediate states. Two-stage training: SFT on CoT-annotated synthetic data + GRPO with verifiable rewards.

Main Evidence: 7.7K synthetic CoT samples outperform 165K real-world Video-R1 samples. RexTime: +12.6% from synthetic training. TOMATO: +4.6% on complex reasoning. 21× data efficiency improvement. Without cold-start SFT, RL training degrades performance (Video-R1 CoT-165K without cold-start: -3.1 to -1.9 on complex reasoning). Short-term training shows largest gains on TemporalBench-action (+4.3%), long-term on TOMATO (+4.1%).

Relevance to Topic: SynRL establishes that abstract temporal primitives transfer from synthetic to real-world video, suggesting that carefully designed training data may be more important than data quantity for temporal understanding.

Limits/Caveats: Training on geometric shapes may not fully capture the complexity of natural video temporal reasoning. The approach requires specific

pipeline design; general applicability to other temporal challenges remains to be demonstrated.

Newly Strengthened / Newly Added Papers from Downloaded PDFs

All 17 papers in the literature collection were successfully refined from downloaded PDFs, strengthening their evidence base with full-text analysis:

1. **VCBench** (2603.12703v2) - PDF extraction confirmed streaming evaluation methodology and detailed metric derivations including GPA, MoC, and UDA formulas.

2. **ReXTime** (2406.19392v2) - PDF provided complete performance tables across 6 frontier models, fine-tuning experiments, and pipeline cost analysis.

3. **MVP** (2506.09987v1) - PDF confirmed language-only baseline results, detailed minimal-pair scoring methodology, and MVBench shortcut analysis.

4. **Video-MME-v2** (2604.05015) - PDF extracted full taxonomy, robustness ratio analysis across model scales, and thinking-mode analysis.

5. **LVBench** (2406.08035v1) - PDF provided ablation studies on LLM filtering effectiveness and per-category performance breakdown.

6. **Neptune** (2412.09582v2) - PDF extracted GEM metric details, frame ablation curves, and Neptune-MMH analysis.

7. **MMOU** (2603.14145v1) - PDF confirmed cross-modal dependency analysis, temporal position sensitivity, and skill-wise breakdown.

8. **BenchScope** (2603.29357v1) - PDF provided ED computation details, weight sensitivity analysis, and ED-Greedy algorithm specifications.

9. **Triage** (2601.22959v1) - PDF extracted hyperparameter sensitivity, per-benchmark performance breakdown, and ablation details.

10. **ContextBudget** (2604.01664v1) - PDF confirmed curriculum stage details, ablations, and RL training configuration.

11. **TimE** (2505.12891v3) - PDF extracted detailed task taxonomy, model comparison tables, and RAG analysis.

12. **TEMPURA** (2505.01583v1) - PDF provided VER dataset construction pipeline details, ablation studies, and qualitative examples.

13. **SynRL** (2603.17693v1) - PDF confirmed per-category transfer patterns, scaling analysis, and cold-start ablation.

14. **TemporalBench** (2410.10818) - Metadata only; PDF extraction unsuccessful for detailed analysis.

15. **TOMATO** (2410.23266) - Limited metadata; PDF extraction unsuccessful.

16. **VideoScore2** (2509.22799) - Generative video evaluation with reasoning-based scoring; PDF extraction unsuccessful.

17. **GRADEO** (2503.02341v1) - Human-like T2V evaluation via multi-step reasoning; PDF extraction unsuccessful.

4.2 Integrated Thematic Assessment

4.2.1 Evaluation Methodology as the Critical Differentiator

The literature reveals that evaluation methodology critically affects conclusions about model capability. Standard per-question accuracy significantly overestimates model capability by 15-25%, as demonstrated by Video-MME-v2's group-based nonlinear scoring that penalizes inconsistent responses. MVP's minimal-change pair design exposes shortcut exploitation that inflated previous benchmarks by 18-20 percentage points. Language-only baselines on MVP achieve 38% versus 25% random baseline, confirming that textual biases contribute to inflated scores. Single-frame baselines on LVBench show minimal performance degradation compared to full-video processing, indicating that temporal information is often not required for correct answers.

The meeting discussion on hybrid evaluation design—combining short rationale/evidence trace generation with constrained answer format—addresses this directly. Rationale scoring against evidence anchors, rather than prose quality, separates genuine reasoning from confident confabulation. This approach is essential for detecting cases where stronger LLMs make outputs sound reasonable while temporal evidence remains wrong.

BenchScope's Effective Dimensionality analysis adds another layer of methodological rigor, revealing 20× variation in measurement breadth across benchmarks with substantial hidden redundancy. Future benchmarks should demonstrate independent signals from existing evaluations through ED analysis before claiming contribution to capability assessment. The ED-Greedy algorithm provides a principled approach for selecting benchmarks that maximize measurement diversity.

4.2.2 Temporal State Maintenance as the Foundational Capability Gap

VCBench's streaming multi-point query design exposes trajectory inconsistencies invisible to single-query benchmarks. The benchmark demonstrates that even state-of-the-art models fail at basic counting tasks when queried at streaming timepoints, with prediction trajectories showing monotonicity violations and delayed responses to state changes. The gap between single-query accuracy and streaming trajectory consistency suggests that current architectures do not maintain persistent world state representations.

The meeting discussion identified state tracking as one of three critical question families, requiring models to maintain what changed, what remained available,

and what is pending throughout video playback. This represents a critical capability gap for applications requiring tracking object identities, event counts, or cumulative information across video duration. VCBench's three complementary metrics—GPA for numerical precision, MoC for trajectory logical self-consistency, and UDA for temporal awareness—provide a framework for measuring this capability that goes beyond single-query accuracy.

Offline evaluation approximates streaming processing through video truncation, potentially overestimating performance since models can re-examine historical content. This is a significant caveat: the evaluation may be optimistic about streaming performance. Future work should develop true streaming evaluation protocols that prevent backward reference.

4.2.3 Long-Context Architectures and Practical Capability

Dedicated long-video models (LLaMA-VID, MovieChat, LWM) perform near-random on LVBench and Neptune, while simpler approaches with more frames outperform them. This suggests fundamental architectural limitations rather than mere processing constraints. Neptune's finding that most benchmarks saturate at ~50 frames while Neptune requires >150 frames indicates that genuine long-context challenge remains unsolved. The literature demonstrates that architectural solutions for long context have not translated to practical capability improvements.

LVBench's results are particularly stark: LLaMA-VID, MovieChat, and LWM are nearly equivalent to random selection on hour-long videos, while LLaVA-NeXT with only 32 input frames outperforms all long-video native models except Gemini 1.5 Pro. This suggests that frame selection quality matters more than the number of frames processed, a finding consistent with Triage's results on hierarchical visual budgeting.

Neptune reveals an additional complication: ASR-heavy content dominates many "long video" datasets, with visual understanding contributing minimally. True long-video benchmarks must emphasize vision-dependent questions to avoid measuring ASR capability rather than visual understanding. The single middle frame outperforms first frame on Neptune, indicating that selecting the right moment is more important than capturing the beginning of the video.

4.2.4 Cross-Segment Reasoning as the Capability Frontier

ReXTime shows the largest human-model gaps (14.3%) for reasoning-across-time tasks where questions and answers occur in different video segments. Moment localization is significantly harder than VQA across all models, with Claude3-Opus achieving only 13.67% accuracy. Cause-effect relationships are particularly challenging among the four event relation types (sequential, cause-effect, means-to-end, and parallel). This represents the capability frontier where architectural innovation is most needed.

The meeting's three question families directly address this: event linking requires connecting moments belonging to the same underlying process, testing whether models can identify that separate video segments describe parts of a continuous causal or sequential chain. This capability is distinct from both simple retrieval (finding a moment) and temporal ordering (identifying sequence).

SynRL demonstrates that 7.7K synthetic samples outperform 165K real-world samples, with RexTime showing +12.6% improvement from synthetic training. This suggests that temporal-centric synthetic data—where answers cannot be inferred from isolated frames—may be particularly valuable for training cross-segment reasoning capabilities. The cold-start SFT finding (RL training degrades without it) has important implications for training pipeline design.

4.2.5 Audio-Visual Integration and Multimodal Benchmarks

MMOU demonstrates that vision-only (44%) and audio-only (35.6% drop) baselines dramatically underperform joint multimodal models (64.2%). Yet most existing benchmarks ignore multimodal integration. LVBench excludes audio data despite providing valuable context, and Neptune-MMH still contains some audio-dependent questions despite filtering. Future benchmarks should require both modalities for correct answers, not merely tolerate them.

The meeting discussion on state tracking emphasizes that audio-visual integration is essential for realistic video understanding. Sound can provide critical evidence for state changes that are difficult to detect visually—doors opening, objects being picked up, liquids being poured. Temporal position sensitivity in MMOU (sharp accuracy drop when evidence appears later in video) compounds with the long-context challenge: even if models can process both modalities, they may fail to integrate information across video duration.

TemporalBench and TOMATO, which could not be fully analyzed due to PDF extraction limitations, remain relevant for understanding fine-grained temporal understanding across action frequency, motion magnitude, and temporal dependency measurement. Their absence from detailed analysis is a gap that future work should address.

4.2.6 Synthetic Data and Training Paradigms

SynRL's finding that 7.7K synthetic samples outperform 165K real-world samples (21× data efficiency) represents a paradigm shift for temporal reasoning training. The key insight is that temporal-centricity and annotation quality matter more than data quantity. Synthetic data design enables precise ground-truth annotation that proprietary model annotations cannot achieve—annotations based on ground-truth trajectories rather than potentially flawed model outputs.

The two-stage training approach (SFT on CoT-annotated synthetic data + GRPO with verifiable rewards) demonstrates that cold-start SFT is essential: without it, RL training degrades performance. This finding has implications for how temporal reasoning capabilities should be trained: first establish basic capability through

supervised learning, then refine through reinforcement learning with verifiable rewards.

TEMPURA's two-stage training (masked event prediction + video segmentation) provides complementary evidence for this approach. Both stages are necessary: masked event prediction alone achieves 35.8 mIoU, segmentation alone achieves 36.1 mIoU, but combined they achieve 39.2 mIoU. This suggests that understanding causal event relationships and fine-grained temporal segmentation are complementary capabilities.

5. Integrated Assessment for the Current Project

The meeting presentation and literature analysis together provide a comprehensive picture of the challenges and opportunities in designing benchmarks for video temporal reasoning. The most justified directions are those that explicitly penalize shortcut behavior, measure consistency rather than just accuracy, and require genuine temporal state maintenance.

Several assumptions are only weakly supported. The assumption that stronger LLMs improve video understanding is complicated by MVP's finding that they can hide grounding failures through plausible-sounding confabulation. The assumption that architectural solutions for long context will translate to capability improvements is contradicted by the near-random performance of dedicated long-video models on LVBench and Neptune. The assumption that more frames always help is contradicted by Triage's results showing that intelligent frame selection at 50% token retention outperforms full processing.

Branches that deserve further testing include: the three question families proposed in the meeting (event linking, state tracking, counterfactual/contrastive), evaluated with minimal-change pairs that penalize shortcut exploitation; the hybrid evaluation approach with rationale scoring against evidence anchors; and the adaptive context budget setting that reflects deployment scenarios.

The current evidence allows us to conclude that benchmark performance significantly overestimates real-world capability, that shortcut exploitation is systematic rather than occasional, that temporal state maintenance is fundamentally underdeveloped, that long-context architectures have not translated to practical capability, and that synthetic data with temporal-centricity can be more effective than large-scale real-world data. The evidence does not yet allow us to conclude which specific architectural innovations will close the human-model gap, how to design benchmarks that remain informative over time, or how to balance fixed-budget and adaptive evaluation settings.

6. Unresolved Questions and Decision-Critical Gaps

1. **How to design benchmarks that remain informative over time**: Popular benchmarks get optimized and stop being informative. The meeting raised the concern that model progress is conflated with annotation design

progress, but no solution was proposed beyond "designing benchmarks as evolving frameworks rather than frozen leaderboards." Specific mechanisms for maintaining benchmark diagnostic value require further development.

2. **Whether retrieval decisions should be part of the benchmark or model design**: The meeting discussion identified this as an unresolved question. If retrieval is part of the benchmark, it can be standardized and fairly compared. If retrieval is part of model design, models can optimize their retrieval strategy, but this conflates two different capabilities. The decision affects what we are actually measuring.

3. **Whether ambiguous cases should be excluded from evaluation**: Excluding ambiguous cases may make the benchmark easier but less representative of real-world use. Including them may make scoring unstable. The meeting discussion did not resolve this tradeoff. This is particularly relevant for counterfactual/contrastive questions where multiple plausible answers may exist.

4. **How to evaluate true streaming performance**: VCBench's offline evaluation approximates streaming through video truncation, potentially overestimating streaming performance since models can re-examine historical content. Developing evaluation protocols that prevent backward reference while maintaining practical feasibility is an open challenge.

5. **Whether synthetic data training transfers to diverse video domains**: SynRL demonstrates temporal primitive transfer from geometric shapes to human activity recognition, but the general applicability to other temporal challenges remains to be demonstrated. Real-world video temporal reasoning may have characteristics that geometric synthetic data cannot capture.

6. **How to balance evaluation rigor with practical accessibility**: Group-based nonlinear scoring (Video-MME-v2) provides more accurate capability estimates but requires more annotation effort and is less intuitive to interpret. Minimal-change pair design (MVP) penalizes shortcut exploitation but may be too strict for some questions where multiple valid answers exist.

7. **The role of audio in video temporal reasoning**: MMOU demonstrates that audio-visual integration is critical, yet most benchmarks exclude audio. Developing benchmarks that require both modalities while maintaining practical annotation costs is an unsolved challenge.

8. **How to measure whether rationale reflects genuine reasoning**: The meeting proposed scoring rationale against evidence anchors rather than prose quality, but implementing this requires defining what counts as sufficient evidence and how to score partial or mixed evidence.

7. Recommended Next Steps

1. **Implement minimal-change pair evaluation for the three question families**: Develop a benchmark variant that requires models to correctly answer both videos in visually similar pairs with opposing answers. This penalizes shortcut exploitation based on superficial visual or textual cues. Use MVP's methodology as a starting point but adapt for the specific question families (event linking, state tracking, counterfactual/contrastive).

2. **Develop streaming evaluation protocol with no backward reference**: Move beyond video truncation approximation. Design evaluation that requires models to maintain predictions at streaming timepoints without access to future content. This addresses VCBench's limitation and provides a more accurate measure of streaming capability.

3. **Implement hybrid rationale scoring with evidence anchor verification**: Combine short rationale generation with constrained answer format. Score rationale against evidence anchors rather than prose quality. This addresses the concern that stronger LLMs can make outputs sound reasonable while temporal evidence remains wrong. Develop protocols for defining sufficient evidence and scoring partial evidence.

4. **Run controlled comparison of synthetic vs. real-world training data**: Following SynRL's methodology, design experiments comparing temporal reasoning improvement from synthetic temporal-centric data versus large-scale real-world data. Focus on the three question families. Measure transfer to diverse video domains beyond the training distribution.

5. **Develop adaptive context budget evaluation**: Implement the two-tier evaluation approach (fixed-budget for fair comparison, adaptive for deployment scenarios). Measure how different model architectures perform under varying context constraints. This addresses the meeting insight that context budget changes the nature of the task.

6. **Investigate audio-visual state tracking**: Develop tasks that require both audio and visual information for state maintenance. MMOU demonstrates that audio-visual integration is critical, yet benchmarks that explicitly measure this capability are lacking. Design annotation protocols that capture audio-visual state changes.

7. **Apply BenchScope methodology to video benchmarks**: Before investing in new benchmark development, analyze existing video benchmarks using Effective Dimensionality metrics. Identify which benchmarks provide independent signals and which are redundant. Use ED-Greedy algorithm to select a diverse evaluation suite.

8. Key Risks, Caveats, and Evidence Boundaries

1. **Evaluation methodology risk**: Standard per-question accuracy significantly overestimates model capability by 15-25%. Conclusions about temporal reasoning capability based on conventional evaluation metrics may be unreliable. Claims of improvement may partially reflect benchmark-specific optimization rather than genuine capability advance.

2. **Shortcut exploitation risk**: MVP demonstrates systematic shortcut exploitation across existing benchmarks. Models may achieve high scores through language priors, single-frame cues, or local visual patterns without genuine temporal understanding. This is not a marginal issue—it affects most existing benchmarks substantially.

3. **Long-context architectural risk**: Dedicated long-video models perform near-random on hour-long videos. The assumption that architectural solutions for long context will translate to practical capability improvements is not supported by current evidence. Simpler approaches with better frame selection may outperform specialized architectures.

4. **Synthetic data transfer risk**: SynRL demonstrates temporal primitive transfer from geometric shapes, but the general applicability to diverse real-world video temporal reasoning is unproven. Training on synthetic data may improve performance on synthetic-like temporal challenges while failing to generalize to natural video.

5. **Benchmark saturation risk**: Popular benchmarks get optimized by the community and stop being informative. Reported model progress may conflate actual capability advance with annotation design progress and benchmark-specific optimization. The evaluation lifecycle creates a moving target for genuine capability assessment.

6. **Audio-visual integration gap**: Most benchmarks exclude audio despite evidence that audio-visual integration is critical for realistic video understanding. Conclusions about video temporal reasoning may not transfer to settings where audio is available and informative.

7. **Grounding verification limitation**: The hybrid evaluation approach (rationale scoring against evidence anchors) is conceptually sound but practically challenging to implement reliably. Defining sufficient evidence and scoring partial evidence requires careful protocol design that has not yet been standardized.

8. **Temporal position sensitivity**: MMOU demonstrates sharp accuracy drops when relevant evidence appears later in videos. Models that perform well on short videos may fail on long videos not because of fundamental capability limitations but because of how evidence is distributed across video duration. This creates a confound in evaluating temporal reasoning capability independent of position effects.