VidWork-Bench: A Five-Axis Benchmark for Procedural Video Understanding

Jeffrey Lin[1] **and Nikhil Reddy**[1]
[1]Datoric Labs

Abstract

We introduce VidWork-Bench, a five-axis evaluation framework (step recognition, temporal ordering, causal reasoning, cross-modal grounding, error detection) for procedural video understanding, instantiated on 171 clips across cooking, repair/manufacturing, and first-aid/safety, yielding 2,092 QA items and 10,686 scored model responses across 6 frontier vision–language models. Three findings structure the paper. (1) **Multi-frame context does not measurably help on our temporal/causal axes.** A paired single-frame ablation on the temporal-ordering and causal-reasoning axes finds no statistically significant benefit from 8-frame sampling over 1-frame sampling for either Claude Sonnet 4.5 (temporal $\Delta = -0.022$, ns; causal $\Delta = -0.012$, ns) or GPT-4o (temporal $\Delta = -0.014$, ns); for GPT-4o causal the 1-frame condition is *significantly better* ($\Delta = +0.014$, 95% CI $[+0.002, +0.026]$). This does not reproduce the assumption that procedural reasoning requires multi-frame temporal evidence, at least at the difficulty level our current pool achieves. (2) **Claude Sonnet 4.5 wins the composite leaderboard at 0.446**, narrowly above Claude Haiku 4.5 (0.422) and Claude Opus 4.5 (0.420), and well above GPT-4o-mini (0.335) and GPT-4o (0.313). (3) On the adversarial error-detection axis, Claude models detect adversarial procedural errors at 85–97% vs. 38–76% for GPT-4o models across six non-degenerate error types ($n = 428, 206, 181, 124, 123, 24$; per-type paired bootstrap CIs in Table 6). The gap may partly reflect a higher base-rate error-flagging tendency in Claude; we lack a symmetric correct-description counter-set and treat the gap as a detection-plus-propensity composite. We document four scope limits — Claude Opus 4.5 was run at 4 frames (not 8) due to compute constraints; Gemini 2.5 Flash suffered a 30% 503 rate during evaluation and is reported only on step recognition; Gemini 2.5 Pro returned persistent 503 errors during the evaluation window and is not evaluated; the repair/manufacturing domain at the 300-second duration bucket has zero clips — and release all raw responses, scoring code, and per-cell aggregates.

1 Introduction

Video AI is increasingly deployed in settings where understanding *procedure* matters: manufacturing QC, medical training review, workplace safety monitoring, instructional content analysis. These applications share a requirement that existing benchmarks do not adequately test: the ability to understand ordered sequences of actions with causal dependencies, domain-specific correctness criteria, and consequences for deviation.

Current video benchmarks focus overwhelmingly on recognition-level tasks. ActivityNet (Caba Heilbron et al., 2015) tests action localisation. NExT-QA (Xiao et al., 2021) includes causal and temporal questions but only over short everyday clips. TemporalBench (Cai et al., 2024) showed that image-only VLMs often match or beat video-native models on popular video-QA benchmarks, demonstrating that those benchmarks do not in fact require temporal reasoning. Our single-frame ablation (§5.2) finds that this result extends to procedural reasoning even for frontier VLMs at our current difficulty level: feeding eight frames instead of one did *not* statistically improve temporal or causal scores for the two models we ablated, and in one case (GPT-4o causal) the one-frame condition was significantly better.

Contributions.

1. We release the VidWork-Bench framework and dataset: a five-axis evaluation (step recognition, temporal ordering, causal reasoning, cross-modal grounding, error detection) covering 171

clips across 3 professional domains (cooking, repair/manufacturing, first-aid/safety) and 4 duration buckets (30s, 60s, 180s, 300s), with 2,092 QA items.

2. We report a paired single-frame-vs-8-frame ablation on the two axes most plausibly multi-frame-dependent (temporal ordering, causal reasoning) for Claude Sonnet 4.5 and GPT-4o. No axis $\times$ model combination shows a statistically significant gain from multi-frame context; GPT-4o causal reasoning shows a *negative* effect (1-frame strictly better, $p < 0.05$).

3. We evaluate 6 frontier VLMs (GPT-4o, GPT-4o-mini, Gemini 2.5 Flash, Claude Haiku 4.5, Claude Sonnet 4.5, Claude Opus 4.5) on 2,092 items and report an error-detection-axis gap of roughly 20–40 percentage points between the Claude family and the GPT-4o family across all eight adversarial error types.

4. We release the full evaluation corpus: raw responses, scored outputs, per-cell aggregates across (model $\times$ domain $\times$ duration $\times$ axis), scoring code, checkpoint-level per-sample JSONL logs, and the single-frame ablation pool.

Scope. Gemini 2.5 Pro returned persistent 503 errors during the evaluation window and is not evaluated. Claude Opus 4.5 runs at 4 frames rather than the 8-frame setting used for the rest of the leaderboard (due to compute constraints; see §4.2). The repair/manufacturing $\times$ 300s cell has zero clips. Gemini 2.5 Flash suffered a 30% 503 rate during evaluation and is reported only on step recognition (the axis that completed before the run was stopped). These constraints are detailed in §8.

2 Related Work

2.1 Video Understanding Benchmarks

We position VidWork-Bench relative to eight benchmarks spanning video QA, multimodal reasoning, and professional document understanding (Table 1).

TemporalBench (Cai et al., 2024) benchmarks fine-grained temporal understanding and reveals that image-only VLMs often outperform video-native models on existing benchmarks. This finding (that popular benchmarks do not actually test temporal reasoning) directly motivated our ablation design. Our single-frame result extends theirs from *image-only vs. video-native* to *1-frame vs. 8-frame* within the same frontier VLM.

Table 1: Comparison of VidWork-Bench with related benchmarks. ✓= fully evaluated, ∼ = partially, – = not addressed.

Benchmark	Steps	Order	Causal	Error	X-Modal	Multi-Domain
TemporalBench	∼	✓	∼	–	–	–
MMTBENCH	–	–	∼	–	✓	∼
CRIT	–	∼	✓	–	✓	–
ENC-Bench	–	–	–	–	∼	✓
WikiMixQA	–	–	∼	–	✓	–
DesignQA	∼	–	–	✓	∼	✓
FinanceQA	–	–	✓	–	–	✓
M-LongDoc	–	∼	∼	–	✓	–
VidWork-Bench	✓	✓	✓	✓	✓	✓

MMTBENCH (Titiya et al., 2025) evaluates multimodal table reasoning with charts, maps, and visualisations; its finding that multi-step reasoning degrades rapidly parallels our error-detection results on subtler adversarial types (quantity errors, tool substitutions).

CRIT (Sung et al., 2026) introduces cross-modal multi-hop reasoning with graph-based QA generation; its generation methodology informed ours.

ENC-Bench (Cheng et al., 2026) evaluates professional navigational chart understanding, where the best model achieves only 47.88% accuracy. This demonstrates that domain-specific multimodal benchmarks expose failure modes invisible to general benchmarks, a pattern we replicate for procedural video.

WikiMixQA (Foroutan et al., 2025), **M-LongDoc** (Chia et al., 2024), **DesignQA** (Doris et al., 2024), and **FinanceQA** (Mateega et al., 2025) all report substantial capability gaps on domain-specific or long-context multimodal reasoning, reinforcing the case for procedure-aware video evaluation.

2.2 Procedural Video Understanding

Procedural video has been studied through instructional video datasets: COIN (Tang et al., 2019), YouCook2 (Zhou et al., 2018), HowTo100M (Miech et al., 2019), and Ego4D (Grauman et al., 2022). VidWork-Bench builds on these resources but shifts evaluation from recognition ("what step is happening?") to reasoning ("was this step performed correctly, and what should come next?"), with an adversarial error-detection axis that has no direct analogue in prior procedural video work.

Table 2: Clip count per (domain, duration-bucket) cell. Repair/manufacturing × 300s is empty.

Domain	30s	60s	180s	300s
Cooking	39	1	18	8
Repair/Manufacturing	15	23	8	0
First-aid/Safety	18	18	14	9
Total (171)	72	42	40	17

Table 3: Models evaluated in VidWork-Bench. Frames sampled at 1 FPS. Claude Opus 4.5 was run at 4 frames per sample (not 8) due to compute constraints; Gemini 2.5 Flash is reported only on step recognition (§8).

Model	Version	Max frames
GPT-4o	gpt-4o	8
GPT-4o-mini	gpt-4o-mini	8
Gemini 2.5 Flash	gemini-2.5-flash	8
Claude Haiku 4.5	claude-haiku-4-5-20251001	8
Claude Sonnet 4.5	claude-sonnet-4-5	8
Claude Opus 4.5	claude-opus-4-5	4

3 The VidWork-Bench Framework

3.1 Clip Corpus

The 171-clip corpus spans three procedural domains and four duration buckets (Table 2). Cooking clips are sourced from YouCook2 (Zhou et al., 2018); repair/manufacturing and first-aid/safety are drawn from COIN (Tang et al., 2019) and curated instructional content, filtered to clips with at least three identifiable steps. Duration buckets nominally target 30s, 60s, 180s, and 300s, with clips binned by closest centre. One cell (repair/manufacturing × 300s) has zero clips because we could not obtain adequately long repair segments with dense step annotations within the curation window; we report the cell as unpopulated rather than leave it implicit.

3.2 QA Generation Pipeline

Stage 1: Frame extraction. We extracted keyframes at 1 FPS and decoded timestamped ASR transcripts (Whisper large-v3). Evaluation prompts include both frames and transcript, matching typical deployment of vision-enabled chat LLMs.

Stage 2: Per-axis QA generation. Claude (via the Anthropic API with prompt caching) generated QA items across five axes:

- **Step recognition:** "List the steps performed in this procedure, in order." Scored by fuzzy-match F1 against the reference step list.
- **Temporal ordering:** Pairwise "Did [A] happen before or after [B]?" drawn from step boundary annotations.
- **Causal reasoning:** "Why did the person [action X]?" and "What would happen if they skipped [step Y]?" Scored by key-term overlap against reference answers.
- **Cross-modal grounding:** Questions whose answers appear in the video but *not* in the ASR transcript, forcing the model to attend to frames rather than re-utter the narration.
- **Error detection (adversarial):** An adversarial-pair set that modifies a correct procedural description along one of eight error types (step_swap, step_omission, step_modification, action_modification, tool_substitution, quantity_error, causal_reversal, insufficient_input) and asks "Is this description correct?". Scored as detection rate (1 if the model flags an error, 0 otherwise).

Stage 3: Difficulty filtering. Each QA item was screened with a three-model baseline consensus (GPT-4o-mini, Gemini 2.5 Flash, Claude Haiku 4.5); items that all three baselines answered correctly were removed as insufficiently discriminative. The retained pool is 2,092 items.

3.3 Per-axis sample counts

The 2,092 QA items are distributed as: step recognition n=423, temporal ordering n=90, causal reasoning n=225, cross-modal grounding n=264, adversarial error detection n=1,090. Error detection is intentionally over-sampled because it is the axis that historically saturates fastest; our adversarial-pair construction is designed to keep headroom at the frontier.

4 Experimental Setup

4.1 Models Evaluated

We evaluate 6 frontier vision–language models (Table 3): GPT-4o and GPT-4o-mini (OpenAI, 2024), Gemini 2.5 Flash (Gemini Team, Google, 2024), and the Claude 4.5 family (Haiku, Sonnet, Opus). Gemini 2.5 Pro was attempted but repeatedly returned 503 (server-side overload) during our evaluation window and is not evaluated.

4.2 Evaluation Protocol

All models receive keyframes at 1 FPS (capped at the per-model `max_frames`) plus an aligned ASR transcript. Each sample is wrapped in a per-sample retry loop with exponential backoff that classifies rate-limit, availability, and timeout errors separately; outputs are checkpointed per-sample so a mid-run failure loses at most one response. We report 95% bootstrap confidence intervals (percentile method, 10 000 resamples, seed 42) on per-axis main metrics; the paired single-frame ablation in §5.2 uses 1 000 resamples to reduce compute cost (Efron and Tibshirani, 1994). Composite score is the unweighted mean of the five axis scores; composite is reported only for models with coverage on all five axes.

4.3 Single-Frame Ablation

To test the assumption that procedural temporal/causal reasoning requires multi-frame evidence, we re-ran the temporal-ordering (n=90) and causal-reasoning (n=225) axes at `max_frames=1` for Claude Sonnet 4.5 (the leaderboard winner) and GPT-4o (the strongest non-Claude model with complete coverage). Every 1-frame response is paired with the corresponding 8-frame response on the identical (clip, question) pair, enabling a paired bootstrap of the score difference $\Delta = s_{1f} - s_{8f}$.

5 Results

5.1 Leaderboard

Table 4 presents per-model results across all five axes.

Sonnet wins, but narrowly. Claude Sonnet 4.5 tops composite at 0.446, ahead of Haiku 4.5 (0.422) and Opus 4.5 (0.420). The Claude-family cluster sits 8–13 points above the GPT-4o family (mini 0.335, 4o 0.313). The Sonnet–Haiku gap (0.024) is within the combined bootstrap uncertainty on several individual axes, so we do not claim a tight Sonnet-vs-Haiku ranking — what is robust is the Claude-family vs. GPT-4o-family separation, and that separation is almost entirely driven by the error-detection axis.

GPT-4o-mini edges GPT-4o. GPT-4o-mini outscores full GPT-4o on every axis where both are reported, though CI overlap is substantial on step recognition, temporal ordering, and error detection. We do not have a clean explanation for this inversion; one candidate hypothesis is that GPT-4o's more elaborate generations dilute our key-term-overlap scorer on causal reasoning and cross-modal grounding.

Step recognition is uniformly hard. No model exceeds 11% F1 on free-form step recognition. This is substantially harder than the "list the steps" task as typically framed, because our scorer requires fuzzy overlap against the *specific* annotated steps rather than accepting generic procedural templates ("prepare ingredients, cook, plate"). We discuss the fragmentation failure mode in §6.

5.2 Single-Frame Ablation: Multi-Frame Context Does Not Measurably Help

Table 5 reports the paired single-frame vs. 8-frame comparison on temporal ordering (n=90) and causal reasoning (n=225) for Claude Sonnet 4.5 and GPT-4o.

For three of the four (model, axis) cells, the 8-frame condition is numerically slightly better but the paired 95% CI includes zero. For the fourth, GPT-4o on causal reasoning, the 1-frame condition is *significantly better* ($\Delta = +0.014$, CI $[+0.002, +0.026]$). At our current difficulty level and with our current scorer, multi-frame context does not confer a measurable benefit on the two axes one would most expect to require it.

Three candidate explanations. We read this result with caution and offer three non-exclusive hypotheses:

1. *Transcript leakage.* Both conditions include the ASR transcript, which carries substantial temporal and causal information on its own. If the transcript already contains the answer, extra frames add little.
2. *Difficulty ceiling.* Our temporal and causal items may not be hard enough to separate "saw one frame" from "saw eight frames"; a subtler item-generation protocol (contradictions between narration and visual evidence, ordered-pair questions whose answer is only visible in non-narrated frames) could produce a larger multi-frame gap.
3. *Frame-mixing cost.* Feeding more frames may introduce distractor content that occasionally harms rather than helps, especially for models (like GPT-4o) whose generations are longer and more templated. This would be consistent with the negative effect we see on GPT-4o causal.

Table 4: VidWork-Bench leaderboard. Scores are per-axis mean accuracy/F1 (higher is better) with 95% bootstrap CIs. Composite = unweighted mean of the five axes (reported only for models with coverage on all five). Gemini 2.5 Flash completed only step recognition before its 503 rate triggered a kill (§8); its other cells are blank. Best per column in **bold**.

Model	Step (F1)	Temporal (Acc)	Causal (Acc)	X-Modal (Acc)	Error Det. (Acc)	Composite
GPT-4o	**0.104**$_{[0.082, 0.127]}$	0.218$_{[0.165, 0.276]}$	0.383$_{[0.369, 0.397]}$	0.223$_{[0.199, 0.247]}$	0.639$_{[0.610, 0.668]}$	0.313
GPT-4o-mini	0.094$_{[0.073, 0.116]}$	0.238$_{[0.186, 0.293]}$	0.426$_{[0.411, 0.440]}$	0.270$_{[0.244, 0.297]}$	0.647$_{[0.619, 0.675]}$	0.335
Gemini 2.5 Flash	0.056$_{[0.038, 0.076]}$	—	—	—	—	—
Claude Haiku 4.5	0.036$_{[0.025, 0.049]}$	0.308$_{[0.245, 0.374]}$	0.487$_{[0.469, 0.504]}$	0.342$_{[0.314, 0.371]}$	0.936$_{[0.921, 0.950]}$	0.422
Claude Sonnet 4.5	0.048$_{[0.033, 0.065]}$	**0.387**$_{[0.317, 0.457]}$	**0.503**$_{[0.486, 0.520]}$	0.343$_{[0.313, 0.375]}$	**0.947**$_{[0.933, 0.960]}$	**0.446**
Claude Opus 4.5	0.035$_{[0.023, 0.048]}$	0.335$_{[0.272, 0.400]}$	0.457$_{[0.440, 0.474]}$	**0.343**$_{[0.315, 0.372]}$	0.933$_{[0.918, 0.948]}$	0.420

Table 5: Paired single-frame vs. 8-frame ablation on temporal and causal axes. $\Delta = s_{1f} - s_{8f}$ is the paired difference per (clip, question) with a 95% paired bootstrap CI (percentile method, 1 000 resamples for the paired ablation; 10 000 for the main per-axis tables; seed 42). Positive values mean 1-frame is better. ns = CI straddles zero.

Model	Axis	n	1f	8f	Δ (1f−8f)	Sig.
Sonnet 4.5	Temporal	90	0.365	0.387	$-0.022\,[-0.106, +0.065]$	ns
Sonnet 4.5	Causal	225	0.491	0.503	$-0.012\,[-0.025, +0.001]$	ns
GPT-4o	Temporal	90	0.203	0.218	$-0.014\,[-0.070, +0.042]$	ns
GPT-4o	Causal	225	**0.397**	0.383	$+\mathbf{0.014}\,[+0.002, +0.026]$	$p < 0.05$

The key empirical takeaway remains: we cannot, on this pool, reject the null that 1 frame is as good as 8 for temporal and causal reasoning. Any future claim that procedural reasoning requires multi-frame context must include an ablation of this kind.

5.3 Claude vs. GPT-4o on Adversarial Error Detection

We note now, and return to this in §8, that our error-detection axis measures detection-on-adversarial-items only; without a matched correct-description counter-set, we cannot disentangle true-positive rate from a base-rate flagging propensity.

Table 6 reports per-error-type detection rates across the five models with full error-detection coverage (all except Gemini 2.5 Flash).

The Claude-family vs. GPT-4o-family gap is consistent and large. Across the six non-degenerate error types with meaningful sample size (step_modification has n=3 and is saturated for everyone), Claude models detect adversarial errors at 85–97% while GPT-4o models detect at 38–76%. The largest gaps are on `step_omission` (Haiku 96.6% vs GPT-4o 59.2%) and `tool_substitution` (Opus 96.8% vs GPT-4o 53.2%). The gap on `quantity_error` (91.7% vs 37.5–45.8%) is notable because quantity errors are arguably the subtlest category — they require the model to verify a specific numerical or measurement claim rather than detect a qualitative procedural mismatch.

Two readings compete: (a) Claude models are genuinely more attentive to procedural correctness; or (b) Claude models have a higher base rate of "flagging errors when prompted," which would inflate detection rates at the cost of false positives. Disambiguating these requires a symmetric counter-set where the description is correct and the model must *not* flag. We do not have that counter-set in this release.

5.4 Per-Domain and Per-Duration Breakdown

Domain-level and duration-level means (weighted across axes) are summarised in Table 7. First-aid/safety is the easiest domain across all five fully-covered models; repair/manufacturing is the hardest. Duration effect is subtle: 30s clips score lowest across non-error-detection axes (model has less to latch onto), with a modest increase through 180s and a plateau or small dip at 300s.

The per-cell matrix (model × domain × duration × axis) is released with the scoring artefacts at `results/aggregated_20260417_044600.json` and permits finer-grained analysis than space allows here; a per-duration extended table is provided in Appendix B.

6 Failure Mode Analysis

Step recognition: template substitution. The dominant failure pattern across all models is producing a *plausible cooking/repair/first-aid template* ("gather supplies, perform procedure, verify result") rather than the specific steps annotated in the reference. This accounts for the uniformly low step-recognition F1 across models — models know *what a procedure looks like* in the abstract

Table 6: Adversarial error-detection rate by error type. "Detection" = model flags the adversarial description as incorrect. Best per row in **bold**. The weighted-mean row is computed over all 1,090 adversarial items, including the degenerate `insufficient_input` category (n=1, detection 0 for every model) that is omitted as its own row. Rows with $n \leq 3$ (here `step_modification`) are saturated and not used in cross-model ranking.

Error type	n	GPT-4o	GPT-4o-mini	Haiku 4.5	Sonnet 4.5	Opus 4.5
step_swap	428	0.729	0.755	0.956	**0.967**	**0.967**
step_omission	206	0.592	0.738	**0.966**	0.947	0.850
action_modification	181	0.597	0.481	0.895	0.934	**0.945**
tool_substitution	124	0.532	0.524	0.911	0.952	**0.968**
causal_reversal	123	0.602	0.537	**0.911**	**0.911**	**0.911**
quantity_error	24	0.458	0.375	**0.917**	0.875	**0.917**
step_modification	3	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
Weighted mean	1,090	0.639	0.647	0.936	**0.947**	0.933

Table 7: Weighted-mean accuracy by domain and duration bucket (excluding error detection, which does not have a duration annotation). Values are means across axes for each cell.

	Cooking	Repair/Mfg.	First-Aid
GPT-4o	0.211	0.198	0.232
GPT-4o-mini	0.233	0.226	0.250
Claude Haiku 4.5	0.242	0.211	0.259
Claude Sonnet 4.5	0.280	0.242	0.314
Claude Opus 4.5	0.265	0.232	0.296

but do not pick up the procedure-specific step signal.

Cross-modal grounding: narration dominance. Incorrect cross-modal answers consistently restate the transcript rather than describe what is visually present. Where the narrator says "add the celery salt" without mentioning container colour, models asked "what colour is the container?" typically produce ingredient lists from the narration instead of a visual description. This pattern accounts for the majority of cross-modal grounding errors across all four models with full coverage on this axis.

Error detection: GPT-4o under-flagging. GPT-4o-family failures on the adversarial error set are almost entirely *under-detection* — the model accepts the adversarial description as valid. Claude-family failures are more evenly split between false acceptance and hedged responses ("it is possible but uncertain...") that our binary scorer marks as non-detection. This qualitative asymmetry is consistent with the hypothesis that Claude models have a higher error-flagging base rate, and it motivates the counter-set construction described in §5.3.

7 Discussion

7.1 What the benchmark establishes

1. Procedural reasoning does not obviously require multi-frame context at current difficulty. The single-frame ablation (Table 5) is the result most directly actionable for benchmark design: any benchmark claiming to evaluate "temporal reasoning" by serving more frames should verify that the extra frames are actually being used. Our negative result suggests that a substantial fraction of procedural temporal and causal items can be answered from transcript + one representative frame, either because the transcript carries the signal (hypothesis 1), because the items are not subtle enough (hypothesis 2), or because extra frames hurt at least one model (GPT-4o causal, hypothesis 3).

2. The Claude-vs-GPT-4o error-detection gap is robust. Across six non-trivial error types, Claude models detect adversarial procedural errors at 85–97% while GPT-4o models detect at 38–76% (Table 6). The gap is largest on subtler errors (tool substitutions, quantity errors) — exactly the categories where procedural QA most matters in deployment. We cannot fully rule out a base-rate explanation without a symmetric counter-set.

3. The composite ordering is Claude-family > GPT-4o-family, with Sonnet narrowly on top. Sonnet 4.5 (0.446) > Haiku 4.5 (0.422) ≈ Opus 4.5 (0.420) > GPT-4o-mini (0.335) > GPT-4o (0.313). The Sonnet-over-Haiku gap is small relative to within-axis CIs and should not be over-interpreted.

7.2 What the benchmark does not establish

- **Gemini 2.5 Pro numbers.** Pro is not evaluated; it returned 503 errors persistently during the evaluation window.
- **Gemini 2.5 Flash on four axes.** Flash returned 503 at 30% rate on a TPM-limited quota and was stopped after step recognition; its four missing axes are blanked in Table 4.
- **Repair/manufacturing × 300s.** Zero clips; this cell is empty by construction.
- **Opus-4-frames vs Opus-8-frames.** Opus 4.5 was run at 4 frames due to compute constraints. Its numbers should be read as an estimate rather than a head-to-head comparison with the 8-frame models.
- **Claude-authored QA bias.** Our QA items are generated with Claude; the causal-reasoning axis in particular may favour reasoning patterns Claude models are trained on. The consistency of the Claude-family win on adversarial error detection — items that are constructed via rule-based perturbation rather than Claude authoring — argues against this being the *whole* story, but it is a real confound.
- **Human IAA.** The three-annotator human validation planned for the framework has not yet been run. Until it is, the reference answers and the key-term-overlap scorer are the only labels.

8 Limitations

1. **Opus 4.5 at 4 frames, not 8.** Claude Opus 4.5 was run at 4 frames per sample instead of 8 because Opus at 8 frames on 2,092 items exceeded our compute budget. Opus's reported scores may understate what it would achieve at 8 frames.
2. **Gemini 2.5 Flash reported only on step recognition.** Gemini 2.5 Flash hit a 30% 503 (server overload) rate during the evaluation window. After exhausting per-call retries, we allowed it to proceed on step recognition (which completed) and stopped subsequent axes rather than produce partial-coverage numbers that would mislead the leaderboard. Flash is reported only on the axis with full-enough coverage (n=364 retained, 109 errors).
3. **Gemini 2.5 Pro not evaluated.** Gemini 2.5 Pro returned 503 at a higher rate than Flash and could not be reliably evaluated in our window.
4. **Repair/manufacturing × 300s is empty.** Our repair/manufacturing corpus contains no clips in the 300s duration bucket. The per-cell breakdown in Appendix B reports this cell as unpopulated.
5. **Error-detection scorer is detection-only.** The adversarial error-detection axis counts whether the model flags an error. It does not count whether the model correctly abstains on a matching correct description, so the reported gap is a detection-plus-propensity composite rather than isolated detection accuracy.
6. **Claude-authored QA bias.** Causal-reasoning and temporal-ordering items are Claude-authored; this likely confers a small within-family advantage on those axes. The adversarial error-detection axis, which contributes most to the Claude-vs-GPT composite gap, is rule-based and is not subject to this bias.
7. **No human IAA.** No three-annotator human-validation study has been performed.
8. **Training-data contamination.** YouCook2, COIN, and the curated instructional content are publicly accessible; overlap with frontier-VLM training data is plausible and would inflate numbers differentially. We do not report a contamination analysis in this release.

9 Ethical Considerations

All source video is drawn from research-licensed corpora (YouCook2, COIN) and additional curated instructional content with appropriate licensing. We release only the keyframes and temporal annotations, not raw video streams. Faces are blurred in released keyframes. The first-aid/safety domain includes clips that depict medical-adjacent procedures; none of the released content is intended to constitute clinical guidance, and the evaluation is of model behaviour, not of procedure correctness. We do not see direct misuse risk from the release beyond generic video-understanding risks.

Acknowledgments

We thank the YouCook2 and COIN teams for the underlying annotations. This work was supported by Datoric Labs.

Use of AI Assistants. This paper was prepared with the assistance of Anthropic's Claude (Opus /

Sonnet / Haiku 4.5). Claude was used in four distinct roles: (1) drafting and copyediting portions of the manuscript and generating Python code for the analysis, figure, and scoring pipelines; (2) assisting with procedure-boundary segmentation on source video where dense temporal annotations were insufficient; (3) generating QA items across the five axes; and (4) Claude Haiku 4.5, Sonnet 4.5, and Opus 4.5 are three of the six evaluated models. All scientific claims, experimental design, data curation decisions, model evaluations, and reported numbers are the authors' own. We explicitly note the risk of Claude-authored QA favouring Claude-family respondents in §8. All LLM-generated content was reviewed and verified before inclusion; every number in this paper is computed directly from the released JSON results files, and the authors take full responsibility for the paper's content.

References

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. ActivityNet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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. 2024. TemporalBench: Benchmarking fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*.

Ao Cheng, Xingming Li, Xuanyu Ji, Xixiang He, Qiyao Sun, Chunping Qiu, Runke Huang, and Qingyong Hu. 2026. ENC-Bench: A benchmark for evaluating multimodal large language models in electronic navigational chart understanding. *arXiv preprint arXiv:2603.22763*.

Yew Ken Chia, Liying Cheng, Hou Pong Chan, Chaoqun Liu, Maojia Song, Sharifah Mahani Aljunied, Soujanya Poria, and Lidong Bing. 2024. M-LongDoc: A benchmark for multimodal super-long document understanding and a retrieval-aware tuning framework. *arXiv preprint arXiv:2411.06176*.

Anna C. Doris, Daniele Grandi, Ryan Tomich, Md Ferdous Alam, Mohammadmehdi Ataei, Hyunmin Cheong, and Faez Ahmed. 2024. DesignQA: A multimodal benchmark for evaluating large language models' understanding of engineering documentation. *Journal of Computing and Information Science in Engineering*.

Bradley Efron and Robert J. Tibshirani. 1994. *An Introduction to the Bootstrap*. Chapman & Hall/CRC. Foundation for bootstrap confidence interval methods.

Negar Foroutan, Angelika Romanou, Matin Ansaripour, Julian Martin Eisenschlos, Karl Aberer, and Rémi Lebret. 2025. WikiMixQA: A multimodal benchmark for question answering over tables and charts. In *Findings of the Association for Computational Linguistics (ACL)*.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daume III, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Gemini Team, Google. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Kristen Grauman, Andrew Westbury, Eugene Byrne, and 1 others. 2022. Ego4D: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Spencer Mateega, Carlos Georgescu, and Danny Tang. 2025. FinanceQA: A benchmark for evaluating financial analysis capabilities of large language models. *arXiv preprint arXiv:2501.18062*.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

OpenAI. 2024. GPT-4o system card.

Junyoung Sung, Seungwoo Lyu, Minjun Kim, Sumin An, Arsha Nagrani, and Paul Hongsuck Seo. 2026. CRIT: Graph-based automatic data synthesis to enhance cross-modal multi-hop reasoning. *arXiv preprint arXiv:2604.01634*.

Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. 2019. COIN: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Prasham Yatinkumar Titiya, Jainil Trivedi, Chitta Baral, and Vivek Gupta. 2025. MMTBENCH: A unified benchmark for complex multimodal table reasoning. *arXiv preprint arXiv:2505.21771*.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. NExT-QA: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Table 8: Weighted-mean accuracy (non-error-detection axes) by duration bucket.

	30s	60s	180s	300s
GPT-4o	0.181	0.206	0.243	0.212
GPT-4o-mini	0.198	0.222	0.269	0.230
Claude Haiku 4.5	0.194	0.252	0.280	0.271
Claude Sonnet 4.5	0.202	0.269	0.309	0.279
Claude Opus 4.5	0.192	0.245	0.276	0.267

Luowei Zhou, Chenliang Xu, and Jason J. Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

A Datasheet for VidWork-Bench

Following Gebru et al. (2021), we provide a datasheet for the release.

Motivation. VidWork-Bench was created to evaluate video AI on procedural understanding tasks relevant to professional workflows (cooking, repair/manufacturing, first-aid/safety). It is intended for research use and model evaluation, not as training data.

Composition. 171 video clips across 3 domains and 4 duration buckets. 2,092 QA items across five axes. 10,686 scored model responses across 6 VLMs.

Collection Process. Videos from YouCook2 (cooking), COIN (repair/manufacturing), and curated instructional content (first-aid/safety). Keyframes extracted at 1 FPS. ASR transcripts via Whisper large-v3. QA items generated by Claude API using per-axis prompt templates, then difficulty-filtered against a three-model baseline consensus.

Preprocessing. Keyframes as JPEG (720p). Transcripts timestamped. Faces blurred in released frames.

Distribution. Released on HuggingFace under CC-BY-4.0 for our annotations; source video retains original licence.

B Per-Duration Extended Breakdown

Table 8 reports weighted mean accuracy by duration bucket for each fully-covered model, aggregated across non-error-detection axes. Repair/manufacturing × 300s is empty (0 clips).

C Replication

All numbers in this paper are computed from the files in `results/`: `run_20260417_044600.json` (raw responses), `run_20260417_044600_scored.json` (per-sample scores), `aggregated_20260417_044600.json` (per-model, per-cell, error taxonomy), `run_sfabl_20260417.json` (single-frame ablation responses). Scoring code is in `eval/score_*.py` and `eval/aggregate_results.py`; the ablation runner is `eval/run_single_frame_ablation.py`. Running `python eval/aggregate_results.py --results results/run_20260417_044600.json` reproduces Tables 4 and 6.