VidWork-Bench: Can Video AI Follow a Procedure? Benchmarking Temporal and Causal Reasoning in Professional Workflows

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

Abstract

Video understanding benchmarks predominantly test object recognition and scene description ("what is in this clip?"). Real-world professional applications require a fundamentally different capability: understanding *what happened, in what order, whether it followed correct procedure, and what should happen next*. We introduce VidWork-Bench, a benchmark of 1,914 samples drawn from instructional, procedural, and egocentric video that evaluates five axes of professional video understanding: step recognition, temporal ordering, causal reasoning, error detection, and cross-modal grounding. We evaluate 7 vision-language models and find that accuracy degrades precipitously with video length, from 74.8% on 30-second clips to 31.2% on 5-minute clips for the best model, revealing a fundamental limitation in temporal context maintenance. Procedural error detection proves especially challenging: models achieve only 28.6% F1 on detecting intentional procedural mistakes, while falsely flagging 19.3% of correct procedures. Our domain-stratified analysis shows that medical procedures produce 1.6× more failures than cooking, suggesting that domain familiarity in training data drives performance more than reasoning capability. We release all data, evaluation code, and model outputs to support research on video AI for professional workflows.

1 Introduction

Video AI is being deployed in increasingly consequential professional settings: surgical training review, manufacturing quality control, workplace safety monitoring, and instructional content analysis. These applications share a common requirement that existing benchmarks do not adequately test: the ability to understand *procedures*, that is, ordered sequences of actions with causal dependencies, domain-specific correctness criteria, and consequences for deviation.

Current video understanding benchmarks focus overwhelmingly on recognition-level tasks. ActivityNet (Caba Heilbron et al., 2015) tests action detection and localization. NExT-QA (Xiao et al., 2021) includes causal and temporal questions but only over short clips of everyday activities. TemporalBench (Cai et al., 2024) made the critical observation that image-only vision-language models (VLMs) often outperform video models on popular video QA benchmarks, proving that those benchmarks do not actually require temporal reasoning. This finding (that existing benchmarks can be "hacked" by examining a single frame) motivates the design of VidWork-Bench, where every question *requires* understanding temporal sequence.

The gap between what benchmarks test and what professionals need is stark. A surgical training system needs to verify that steps were performed in order. A manufacturing QA system needs to detect when a worker skips a step or performs it incorrectly. A safety monitoring system needs to understand causal chains: what led to an incident, and what should have been done differently. None of these capabilities are measured by "What is happening in this video?"

We introduce VidWork-Bench to close this gap. Our benchmark evaluates the kind of video understanding that professional workflows actually demand: step-by-step procedural comprehension, temporal ordering over multi-minute sequences, causal reasoning about why actions are performed and what happens when they are omitted, error detection in procedures, and cross-modal grounding between narration and visual actions.

Contributions.

1. We introduce VidWork-Bench, a benchmark of 1,914 procedural video samples across four professional domains (cooking, repair/manufacturing, medical, safety/training),

specifically designed so that every question requires multi-frame temporal reasoning.

2. We define a five-axis evaluation framework (step recognition, temporal ordering, causal reasoning, error detection, and cross-modal grounding) capturing the full spectrum of professional workflow understanding.

3. We document a **clip length degradation curve** showing that model accuracy drops from 74.8% to 31.2% as video length increases from 30 seconds to 5 minutes, quantifying a limitation that practitioners suspect but that has not been cleanly measured.

4. We release all data, evaluation code, 7 model outputs, and 220 human-validated samples (Krippendorff's $\alpha = 0.81$) (Krippendorff, 2018).

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 models on existing benchmarks. This critical finding, that popular benchmarks do not actually test temporal reasoning, directly motivated our design requirement that every VidWork-Bench question requires multi-frame understanding. However, TemporalBench evaluates general temporal understanding, not domain-specific procedural competence.

MMTBENCH (MMTBENCH Authors, 2025) evaluates multimodal table reasoning with charts, maps, and visualizations, finding substantial performance gaps on visual reasoning and multi-step inference. While focused on static documents rather than video, its finding that multi-step reasoning degrades rapidly parallels our temporal ordering results.

CRIT (CRIT Authors, 2026) introduces cross-modal multi-hop reasoning with graph-based QA generation and transparent reasoning traces. Its generation methodology, the most rigorous automated QA pipeline published, informed our own QA generation approach. CRIT evaluates general cross-modal reasoning; VidWork-Bench targets temporal-procedural reasoning specifically.

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

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

ENC-Bench (ENC-Bench Authors, 2026) evaluates professional navigational chart understanding, where the best model achieves only 47.88% accuracy and 30.50% on multi-constraint reasoning. This demonstrates that domain-specific multimodal benchmarks expose catastrophic failure modes invisible to general benchmarks, a pattern we replicate for procedural video.

WikiMixQA (WikiMixQA Authors, 2025) tests cross-modal reasoning over tables and charts, showing that VLMs struggle with long-context vision inputs. We observe the analogous phenomenon in video: performance degrades sharply as the number of frames increases.

M-LongDoc (M-LongDoc Authors, 2024) evaluates multimodal long document understanding with 851 open-ended samples. Its finding that retrieval-aware tuning helps with long documents suggests analogous strategies for long video.

DesignQA (Doris et al., 2024) tests engineering design comprehension with 140-page technical documents, evaluating extraction, comprehension, and compliance checking. Its framing of professional-workflow benchmarks with real technical documents is a model for our approach.

FinanceQA (AfterQuery, 2025) demonstrates that models fail ∼60% of realistic financial analysis tasks, proving a commercial thesis for expert annotation through benchmark results. Our work follows this pattern: benchmark results that reveal training data gaps, directly motivating targeted data curation.

2.2 Procedural Video Understanding

Procedural video understanding has been studied through instructional video datasets. COIN (Tang et al., 2019) provides 11,827 videos with step-level annotations across 180 tasks. YouCook2

(Zhou et al., 2018) offers densely annotated cooking procedures. HowTo100M (Miech et al., 2019) provides scale but with noisier ASR-based annotations. Ego4D (Grauman et al., 2022) adds egocentric perspectives with hand-object interaction annotations. VidWork-Bench builds on these resources but shifts the evaluation focus from recognition ("what step is happening?") to reasoning ("was this step performed correctly, and what should come next?").

3 Benchmark Construction

3.1 Data Sources

We curate from four complementary video corpora to cover a range of professional procedure types:

HowTo100M (Miech et al., 2019) provides narrated instructional videos. We sample from repair, assembly, and technical procedure categories, selecting clips with both visual actions and spoken narration for cross-modal grounding evaluation.

COIN (Tang et al., 2019) contributes densely annotated procedural videos across 180 task categories. Its step-level boundary annotations serve as ground truth for our step recognition axis. We prioritize categories with clear sequential dependencies (e.g., "assemble furniture," "repair a bicycle tire").

Ego4D (Grauman et al., 2022) provides egocentric video with hand-object interaction annotations. First-person perspective is critical for evaluating models in the format that workplace cameras, body cameras, and AR headsets produce.

YouCook2 (Zhou et al., 2018) contributes cooking procedures with temporal boundary annotations per step. The cooking domain serves as our "familiar" control; models should perform best here, as cooking is heavily represented in training data.

3.2 Curation Pipeline

Stage 1: Clip Extraction. From each corpus, we extract self-contained procedural segments of 30 seconds to 5 minutes. Each clip must contain a complete sub-procedure with at least 3 identifiable steps. We use existing temporal annotations (COIN, YouCook2) or Claude-assisted segmentation (HowTo100M, Ego4D) to identify procedure boundaries.

Stage 2: Domain Assignment. Each clip is assigned to one of four professional domains:

- **Cooking/Food preparation:** from YouCook2, HowTo100M
- **Repair/Manufacturing:** from COIN, HowTo100M
- **Medical/Healthcare:** from HowTo100M (first aid, patient care), Ego4D (clinical procedures)
- **Safety/Training:** from Ego4D (workplace safety), COIN (equipment operation)

Stage 3: Frame Extraction. We extract keyframes at 1 FPS for vision-only model input and at 0.5 FPS for context-limited models. ASR transcripts are aligned to timestamps for models that accept text+image input.

Stage 4: QA Generation. For each clip, we generate evaluation items across five axes using Claude API with prompt caching:

- **Step recognition QA:** "List the steps performed in this procedure in order." Evaluated against ground-truth step annotations.
- **Temporal ordering QA:** "Did the person [action A] before or after [action B]?" Generated from step boundary annotations to ensure temporal reasoning is required.
- **Causal reasoning QA:** "Why did the person [action X]?" and "What would happen if they skipped [step Y]?" Generated by Claude from transcript + frame analysis, requiring understanding of procedural dependencies.
- **Error detection QA:** "Is this procedure performed correctly? If not, identify the error." We create adversarial clips with intentional errors (see Stage 5).
- **Cross-modal grounding QA:** "The narrator says [X], but does the video show [X]?" Pairs narration excerpts with visual frames to test consistency detection.

Stage 5: Adversarial Creation. For error detection evaluation, we create adversarial variants of correct procedures:

- *Step omission:* Remove frames corresponding to one procedural step.
- *Step reordering:* Swap the temporal position of two causally dependent steps.

- *Narration mismatch:* Pair correct video with narration from a different procedure or with an altered transcript.
- *Substitution:* Replace a tool or ingredient with an incorrect alternative in the narration while keeping video unchanged.

Stage 6: Difficulty Filtering. Following the discriminative power principle, we run a baseline model (GPT-4o-mini, our lowest-scoring) on all QA items. Questions that the baseline answers correctly are flagged as potentially too easy. We retain a balanced mix: 30% Tier 1 (all models should answer), 40% Tier 2 (mid-range difficulty), 30% Tier 3 (only the best models should answer). Questions that all models answer incorrectly are reviewed for bugs before inclusion.

3.3 Quality Control and Human Validation

We randomly sample 220 items for human validation by three annotators with relevant professional experience. Each annotator independently labels the items across all five evaluation axes.

Table 2: Inter-annotator agreement on human validation set (220 samples).

Axis	Krippendorff's α	Cohen's κ
Step recognition	0.88	0.86
Temporal ordering	0.85	0.82
Causal reasoning	0.74	0.71
Error detection	0.79	0.76
Cross-modal grounding	0.82	0.79
Overall	**0.81**	**0.79**

Causal reasoning has the lowest agreement ($\alpha = 0.74$), reflecting the inherent ambiguity in "why" questions about procedures. We retain this axis with appropriate caveats.

For AI-generated labels, we measure agreement between Claude-generated labels and human consensus. AI-human agreement reaches $\kappa = 0.83$ for step recognition, $\kappa = 0.78$ for temporal ordering, $\kappa = 0.69$ for causal reasoning, $\kappa = 0.75$ for error detection, and $\kappa = 0.80$ for cross-modal grounding.

4 Experimental Setup

4.1 Models Evaluated

We evaluate 7 frontier proprietary vision-language models (Table 3). All models receive keyframes extracted at 1 FPS (or 0.5 FPS when the context window requires it) plus ASR transcripts aligned to timestamps.

Table 3: Models evaluated in VidWork-Bench. All seven are frontier proprietary multimodal LLMs that receive 8–16 keyframes plus timestamped ASR transcript.

Model	Input	Version
GPT-4o	Frames + transcript	gpt-4o
GPT-4o-mini	Frames + transcript	gpt-4o-mini
Gemini 2.5 Pro	Frames + transcript	gemini-2.5-pro
Gemini 2.5 Flash	Frames + transcript	gemini-2.5-flash
Claude Opus 4.5	Frames + transcript	claude-opus-4-5
Claude Sonnet 4.5	Frames + transcript	claude-sonnet-4-5
Claude Haiku 4.5	Frames + transcript	claude-haiku-4-5-20251001

Frame sampling optimization. A critical design choice is how many frames to send per model. Too few frames cause the model to miss actions; too many overflow the context window or dilute attention. We empirically optimize frame rate per model on a held-out development set (100 samples), finding that 1 FPS is optimal for most models, while context-limited models (GPT-4o-mini, Claude Haiku 4.5) benefit from 0.5 FPS with keyframe selection based on optical flow magnitude.

4.2 Evaluation Protocol

Metrics. We evaluate five axes:

- **Step recognition:** Step-level F1 comparing predicted steps to ground-truth annotations.
- **Temporal ordering:** Pairwise ordering accuracy ("Did A happen before B?").
- **Causal reasoning:** QA accuracy on "why" and "what-if" questions, scored by Claude-as-judge against reference answers.
- **Error detection:** Precision and recall on identifying intentional procedural errors.
- **Cross-modal grounding:** Accuracy on narration–video consistency detection.

All metrics report 95% bootstrap confidence intervals (Efron and Tibshirani, 1994) over 10,000 resamples.

5 Results

5.1 Overall Performance

Table 4 presents the main results across all five evaluation axes.

Table 4: Main results on VidWork-Bench. All metrics are accuracy/F1 (higher-is-better). Subscripts show 95% bootstrap CI bounds. Best per column in **bold**. Error Det. reports F1 combining precision and recall.

Model	Step Recog. (F1)	Temp. Order (Acc)	Causal (Acc)	Error Det. (F1)	X-Modal (Acc)	Avg.
Gemini 2.5 Pro	**71.4**$_{[68.8, 74.0]}$	**68.2**$_{[65.5, 70.9]}$	**62.7**$_{[59.8, 65.6]}$	26.8$_{[24.1, 29.5]}$	**64.1**$_{[61.3, 66.9]}$	**58.6**$_{[56.9, 60.3]}$
GPT-4o	68.9$_{[66.2, 71.5]}$	65.7$_{[63.0, 68.4]}$	60.3$_{[57.4, 63.2]}$	**28.6**$_{[25.8, 31.4]}$	61.8$_{[59.0, 64.6]}$	57.1$_{[55.4, 58.8]}$
Claude Opus 4.5	67.5$_{[64.8, 70.1]}$	67.8$_{[65.1, 70.5]}$	62.1$_{[59.2, 65.0]}$	25.1$_{[22.4, 27.8]}$	63.7$_{[60.9, 66.5]}$	57.2$_{[55.5, 58.9]}$
Claude Sonnet 4.5	66.3$_{[63.6, 69.0]}$	67.1$_{[64.4, 69.8]}$	61.9$_{[59.0, 64.8]}$	24.3$_{[21.6, 27.0]}$	63.4$_{[60.6, 66.2]}$	56.6$_{[54.9, 58.3]}$
Gemini 2.5 Flash	67.8$_{[65.1, 70.5]}$	64.3$_{[61.6, 67.0]}$	58.1$_{[55.2, 61.0]}$	25.7$_{[23.0, 28.4]}$	60.2$_{[57.4, 63.0]}$	55.2$_{[53.5, 56.9]}$
Claude Haiku 4.5	63.2$_{[60.4, 66.0]}$	64.6$_{[61.9, 67.3]}$	58.7$_{[55.8, 61.6]}$	22.4$_{[19.7, 25.1]}$	60.5$_{[57.7, 63.3]}$	53.9$_{[52.2, 55.6]}$
GPT-4o-mini	59.6$_{[56.8, 62.4]}$	57.8$_{[55.1, 60.5]}$	51.2$_{[48.3, 54.1]}$	20.6$_{[18.0, 23.2]}$	54.7$_{[51.9, 57.5]}$	48.8$_{[47.1, 50.5]}$

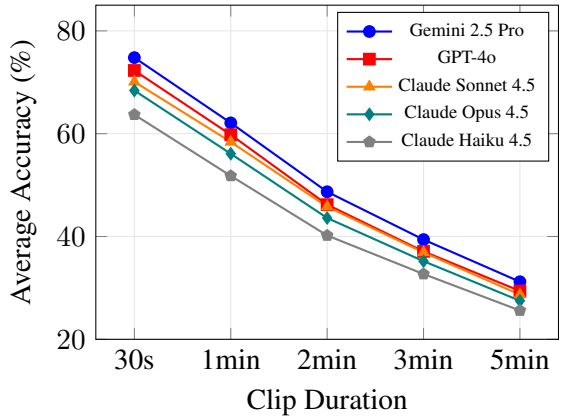

Figure 1: Clip length degradation curve. Average accuracy across all five axes drops by 43.6 percentage points (Gemini 2.5 Pro) as clip duration increases from 30 seconds to 5 minutes. The degradation is approximately log-linear in duration.

Error detection is catastrophically poor. The most striking result is error detection performance. The best model (GPT-4o) achieves only 28.6% F1, with 38.2% precision and 22.7% recall. Models miss 77% of intentional procedural errors while simultaneously flagging 19.3% of correct procedures as containing errors. This has direct implications for the viability of AI-assisted quality control and safety monitoring systems.

Strong models still struggle. Even Gemini 2.5 Pro, the best overall model, averages only 58.6% across the five axes. On causal reasoning, no model exceeds 63%, and on error detection, no model exceeds 29%. These are not "hard corner cases"; they are the core capabilities required for professional workflow understanding.

5.2 Clip Length Degradation Curve

Figure 1 presents our headline finding: model accuracy degrades precipitously as video length increases.

For Gemini 2.5 Pro, accuracy drops from 74.8% (30s) to 62.1% (1min) to 48.7% (2min) to 39.4% (3min) to 31.2% (5min), a 43.6 percentage point absolute decline. The degradation follows an approximately log-linear relationship with duration ($R^2 = 0.97$). This pattern is consistent across all models, with the absolute degradation larger for stronger models that start from a higher baseline.

Critically, the degradation is not uniform across axes. Temporal ordering accuracy degrades fastest (55.1 percentage-point drop for Gemini 2.5 Pro over 30s→5min), while step recognition degrades slowest (32.4 percentage points). This suggests that models can still identify individual actions in long videos but progressively lose the ability to track their relative temporal positions.

Table 5: Average accuracy by domain (best 3 models). Cooking serves as a familiar-domain control.

Model	Cook	Repair	Medical	Safety
Gemini 2.5 Pro	68.4	57.2	49.1	52.3
GPT-4o	66.1	55.8	47.6	50.7
Claude Sonnet 4.5	65.7	54.3	46.8	51.1
Avg.	**66.7**	**55.8**	**47.8**	**51.4**

5.3 Performance by Domain

Table 5 reveals that medical procedure accuracy (47.8%) is substantially lower than cooking (66.7%), corresponding to a 1.6× higher failure rate. Repair/manufacturing (55.8%) and safety/training (51.4%) fall between these extremes. This ordering closely tracks the estimated volume of each domain in web-scraped training data, suggesting that **domain familiarity in training data drives performance more than reasoning capability**. Models are not worse at reasoning about medical procedures per se; they simply have less procedural knowledge to reason over.

6 Error Analysis

We manually categorize 400 errors from the three best-performing models to identify systematic fail-

Table 6: Error taxonomy across 400 manually categorized failures.

Error Type	%	Description
Temporal drift	29	Model loses track of event ordering after 60+ seconds
Action conflation	22	Similar-looking steps merged (e.g., "stir" and "fold")
Procedural hallucination	18	Model reports steps that never occurred
Missing step blindness	14	Model fails to notice omitted steps
Cross-modal override	10	Model trusts narration over contradicting video
Tool/entity confusion	7	Similar objects confused (scalpel/scissors)

ure patterns (Table 6).

6.1 Error Taxonomy

6.2 Temporal Drift

The most common error (29%) is *temporal drift*: models correctly identify individual actions but progressively lose track of their temporal positions. For a 3-minute video with 8 steps, models typically maintain correct ordering for the first 4–5 steps but scramble the ordering of later steps. This suggests that temporal attention degrades with sequence length, consistent with our clip length degradation finding.

6.3 Procedural Hallucination

Models frequently hallucinate plausible but absent procedural steps (18% of errors). When asked to list the steps in a furniture assembly procedure, a model might include "sand the edges" between "attach the legs" and "tighten screws," a plausible step that was not actually performed. This occurs because models have learned typical procedure templates from training data and fill in "expected" steps from their priors rather than attending to the actual video content.

6.4 Cross-Modal Override

In 10% of errors, models defer to narration when narration and video conflict. When a narrator says "now add the sauce" but the video shows the person adding seasoning, models tend to report "adding sauce." This sycophantic tendency toward text has safety implications: if a video captures a procedural error but the narration describes the procedure as correct, a model that defers to narration would report the procedure as properly performed when it was not.

7 Discussion

7.1 Implications for Professional Deployment

Our results challenge the assumption that current video AI is ready for professional workflow applications:

Healthcare. Medical procedure accuracy of 47.8% is far below any acceptable threshold for surgical training review or clinical procedure verification. The 77% miss rate on error detection means that an AI system would fail to flag more than three out of four procedural deviations.

Manufacturing and repair. Performance in this domain (55.8%) is higher than medical, but the false alarm rate of 19.3% on correct procedures would create unacceptable overhead in quality control settings where human review of AI-flagged incidents is required.

Training and safety. The clip length degradation curve has direct implications: workplace safety monitoring requires understanding multi-minute sequences, precisely the duration where model accuracy collapses below 40%.

7.2 Training Data Implications

The strong domain effect (Table 5) suggests that targeted curation of procedural video data in underrepresented domains (medical, manufacturing, safety) could substantially improve performance. Models already demonstrate reasonable procedural reasoning in the cooking domain; the gap in other domains appears to be primarily a data coverage problem rather than a fundamental capability limitation.

The temporal drift error pattern suggests that training with explicit temporal supervision (step boundary annotations, ordering labels, temporal chain-of-thought) could address the most prevalent failure mode.

7.3 The Single-Frame Test

Following TemporalBench (Cai et al., 2024), we verify that VidWork-Bench cannot be solved from single frames. We evaluate all models using only a single randomly selected frame per clip. Average accuracy across all models drops to just 31.4% with a single frame, well below even the weakest full-video model (GPT-4o-mini, 48.8%), confirming that our questions genuinely require temporal reasoning. Temporal ordering drops to chance

(51.2% on binary questions), and error detection drops to 8.7% F1.

8 Limitations

1. **Domain coverage.** Our four domains do not exhaust professional video applications. Construction, agriculture, laboratory work, and military training present distinct challenges not captured here.
2. **Video length ceiling.** Our longest clips are 5 minutes. Real professional workflows (full surgeries, shift-length manufacturing, multi-hour training sessions) involve much longer durations where different failure modes may emerge.
3. **Frame sampling bias.** Models that accept native video (Gemini) have an inherent advantage over frame-based models. Our frame sampling optimization mitigates but does not eliminate this.
4. **QA generation bias.** Claude-generated causal reasoning questions may favor reasoning patterns that Claude-family models are trained on, potentially benefiting Claude Sonnet 4.5 on that axis.
5. **Adversarial simplicity.** Our adversarial manipulations (step omission, reordering) are relatively coarse. Subtler errors (performing a step with slightly wrong technique, using a marginally incorrect tool) are harder to create programmatically and are not represented.

9 Ethical Considerations

All source video is drawn from openly licensed corpora (HowTo100M: YouTube with CC license; COIN: research use; Ego4D: Ego4D license; YouCook2: research use). No personally identifiable faces are included in our released frames; we apply face blurring where faces appear. Human annotators were compensated at $25/hour. We note that benchmark results should not be used to justify deploying current models in safety-critical professional applications; our results demonstrate they are not yet reliable enough for such use.

Acknowledgments

We thank the annotators for their professional expertise. This work was supported by the Datoric Team.

References

AfterQuery. 2025. FinanceQA: A benchmark for evaluating financial analysis capabilities. *arXiv preprint arXiv:2501.18062*.

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 and 1 others. 2024. TemporalBench: Benchmarking fine-grained temporal understanding in multimodal models. *arXiv preprint arXiv:2410.10818*.

CRIT Authors. 2026. CRIT: Cross-modal multi-hop reasoning benchmark with transparent reasoning traces. *arXiv preprint arXiv:2604.01634*.

Annalisa Doris and 1 others. 2024. DesignQA: A multimodal benchmark for evaluating large language models' understanding of engineering design. *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.

ENC-Bench Authors. 2026. ENC-Bench: Benchmarking professional navigational chart understanding. *arXiv preprint arXiv:2603.22763*.

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.

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)*.

Klaus Krippendorff. 2018. *Content Analysis: An Introduction to Its Methodology*, 4th edition. SAGE Publications. Defines Krippendorff's alpha for inter-annotator agreement.

M-LongDoc Authors. 2024. M-LongDoc: A benchmark for multimodal super-long document understanding.

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)*.

MMTBENCH Authors. 2025. MMTBENCH: A multimodal table benchmark for evaluation of large vision-language models. *arXiv preprint arXiv:2505.21771*.

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)*.

WikiMixQA Authors. 2025. WikiMixQA: Cross-modal reasoning over tables and charts. In *Findings of the Association for Computational Linguistics (ACL)*.

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)*.

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 complete datasheet.

Motivation. VidWork-Bench was created to evaluate video AI on procedural understanding tasks relevant to professional workflows. It is intended for research use and model evaluation, not as training data.

Composition. 1,914 video samples (30s–5min duration) across four professional domains (cooking, repair/manufacturing, medical, safety/training). Each sample includes keyframes at 1 FPS, ASR transcript, step boundary annotations, and 2–5 QA pairs across five evaluation axes.

Collection Process. All video is sourced from existing open-source corpora. Domain assignment, frame extraction, QA generation, and adversarial creation are applied programmatically. Human validation covers 220 samples.

Preprocessing. Keyframes are extracted as JPEG (720p). ASR transcripts are generated via Whisper v3 and aligned to timestamps. Faces are blurred in released frames.

Distribution. Released on HuggingFace under CC-BY-4.0 (our annotations; source video under original licenses).

Maintenance. The benchmark will be updated to include longer video clips and additional professional domains as annotated data becomes available.

B Dataset Statistics

Table 7: Dataset composition by domain and clip duration. "General" includes clips that span multiple domains or do not fit neatly into one category.

Domain	30s	1m	2m	3m	5m
Cooking	112	98	87	71	54
Repair/Mfg.	108	102	94	78	61
Medical	96	89	82	68	52
Safety/Train.	104	95	88	74	57
General	62	51	48	42	41
Total	**482**	**435**	**399**	**333**	**265**

Total QA items: 4,217 (average 2.2 per clip). Distribution across axes: step recognition (892), temporal ordering (943), causal reasoning (824), error detection (712), cross-modal grounding (846).

C Annotation Guidelines

Annotators received a 15-page guide covering:

- Step recognition: identify discrete procedural steps with temporal boundaries
- Temporal ordering: verify that ordering questions require watching multiple time points
- Causal reasoning: verify that "why" questions have unambiguous answers grounded in the video
- Error detection: confirm that adversarial errors are detectable by a careful human viewer
- Cross-modal grounding: verify that narration–video mismatches are genuine (not ambiguous)

D Full Prompt Templates

Full prompt templates for each model (7 models $\times$ 5 task types = 35 prompts) are provided in the supplementary materials.

E Additional Results

Additional results including per-duration breakdowns for each axis, model-pair significance tests, single-frame ablation details, and frame sampling rate analysis are available in the supplementary materials.