A zero-pixel clock recovers most of the surgical-phase signal on Cholec80

Muhammad Ahmed Cheema*[1], Zaigham Randhawa†[2], and Truffle[3]

[1] *Ghostwright*
[2] *Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV*
[3] *Truffle Discovery Lab*

2026-06-21

Truffle Discovery Lab, Discovery No. 008 • Surgical video

Abstract

Cholec80 is one of the most-cited benchmarks in surgical computer vision, and every model on its leaderboard looks at the video. We ask how much of the reported phase-recognition accuracy is recoverable from the surgical schedule alone. A content-free predictor that receives only a frame's normalized position in time, with no pixels, reaches 70.9% plus or minus 0.6% video-level phase accuracy on matched 40/40 splits, far above the 39.6% majority floor and within reach of PhaseNet's published 78.8%, closing 62% of the floor-to-ceiling gap. The clock is phase-specific: it nails the long, stereotyped bookend phases and is blind to the short interior phases that genuinely require the camera. We also show the official EndoNet split is not exchangeable: the clock scores 19.2 points lower on it than on matched random splits because the two halves come from different duration and phase-mix distributions, so part of every model's reported error on the standard split is a composition mismatch baked into the partition rather than a modeling failure. A label-shuffling control collapses both the floor and the clock to the majority floor, confirming the exploited signal is genuine surgical structure. The result is a reusable, content-free measuring stick for how much of any reported number is the schedule.

Dataset. Cholec80. **Question.** How much of the phase-recognition accuracy that vision models report on Cholec80 is recoverable from the clock alone, with no pixels? **phase accuracy with zero pixels (matched splits):** 70.9%.

The question

Cholec80 is one of the most-cited benchmarks in surgical computer vision. Eighty laparoscopic cholecystectomy videos, each labeled frame by frame with one of seven surgical phases, and a decade of models climbing a leaderboard from the high 70s into the low 90s. Every one of those models looks at the video.

But surgery has a schedule. A cholecystectomy almost always runs in the same order: preparation, then the long Calot-triangle dissection, then clipping and cutting, then the gallbladder comes off

*cheemawrites@gmail.com
†zar00002@mix.wvu.edu

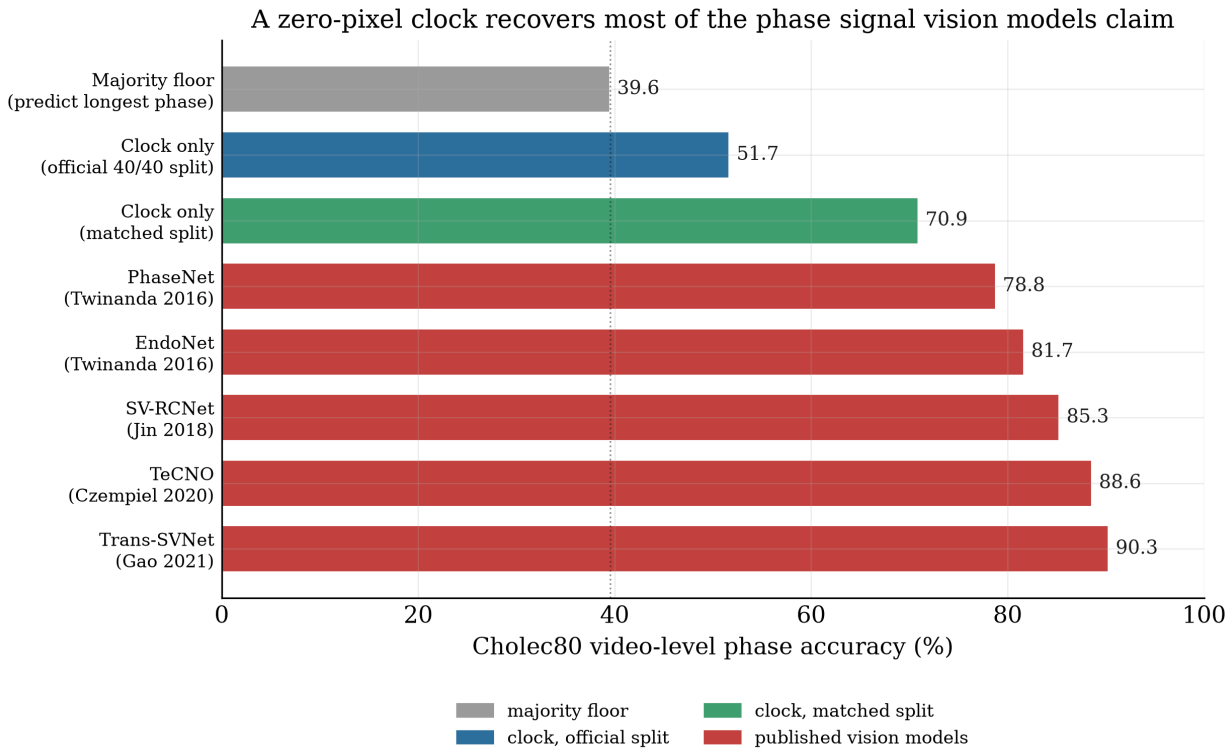

Figure 1: The clock alone (matched split) lands at 70.9%, far above the 39.6% majority floor and within reach of PhaseNet's published 78.8%. None of the published vision models see less than the clock does.

the liver bed, gets bagged, the field is cleaned, and the specimen is retrieved. If the order is that stereotyped, a model might score well by learning *when* things happen rather than *what* it sees.

So I built the most honest version of that shortcut and measured it. A **content-free** predictor that receives, for each frame, only its normalized position in time. No pixels. No instruments. No motion. Just a clock. How much of the benchmark does the clock alone explain?

The novelty gate

Before any modeling, I checked whether this had been done. There is a literature on temporal priors for phase recognition (positional encodings, timestamp supervision, metric critiques), but no published work reports a *vision-free, time-position-only* baseline on the standard Cholec80 phase benchmark. The field has critiqued the metrics; it has not measured the shortcut. That gap is the study.

What I did

Four predictors, each given strictly less information than a vision model:

- **Majority floor.** Always predict the single longest phase. This is the floor any model must clear to claim it learned anything.

- **Clock (histogram).** A transparent 100-bin histogram of normalized time, argmax phase.
- **Clock (gradient-boosted).** A small gradient-boosted tree on normalized time.
- **Elapsed-minutes.** The online variant, given only absolute minutes since the start with no knowledge of total duration.

And three protocols. The official EndoNet split (train videos 1-40, test 41-80). Twenty repeated random 40/40 splits, for a composition-balanced estimate. And a negative control: shuffle each video's phase labels in time, so the clock can carry no real phase information and every model must collapse back to the floor.

The finding

One. A zero-pixel clock recovers most of the phase signal.

With no visual input at all, the clock reaches **70.9% plus or minus 0.6%** video-level accuracy on matched splits, and **51.7%** on the official split. The majority floor is 39.6% and the best published vision model, Trans-SVNet, reports 90.3%. So the clock alone closes **62%** of the floor-to-ceiling gap, and on its own nearly matches PhaseNet's published 78.8%, the model that put this benchmark on the map.

That does not make the clock a good phase recognizer. It is phase-specific in a way that is itself the point: it nails the long, stereotyped bookend phases (Calot dissection F1 0.69, retrieval 0.62, cleaning 0.51) and is completely blind to the short interior phases that genuinely require the camera (clipping and cutting F1 0.00, packaging 0.11). The honest reading is that a meaningful share of what Cholec80 vision models are credited for is the canonical surgical schedule, not scene understanding, and any model that does not clear the clock-only number on a given phase has demonstrated no visual understanding of that phase. The clock is a far stronger floor than the majority class, and it is the floor that matters.

Two. The official Cholec80 split is not exchangeable.

The clock scores 19.2 points *lower* on the official split than on matched random splits. That gap is not noise. The two halves of the standard EndoNet split come from different distributions: training surgeries average 3468 seconds, test surgeries only 2262, with a correspondingly different phase-time composition. Part of every model's reported error on the standard split is therefore a composition mismatch baked into the partition, not a modeling failure. The benchmark looks harder than the task is.

The control holds. Time-shuffling the labels collapses both the floor and the gradient-boosted clock to exactly 39.6%, the majority floor. The temporal signal the clock exploits is genuine surgical structure, not an artifact of the experiment.

Why it matters

This is not a takedown of vision models for surgical phase recognition. They clear the clock, and the interior-phase numbers show they have to be doing real visual work to do it. The contribution is a measuring stick the field was missing: a content-free floor that says how much of any reported number is the schedule, plus a concrete demonstration that the canonical split conflates model error

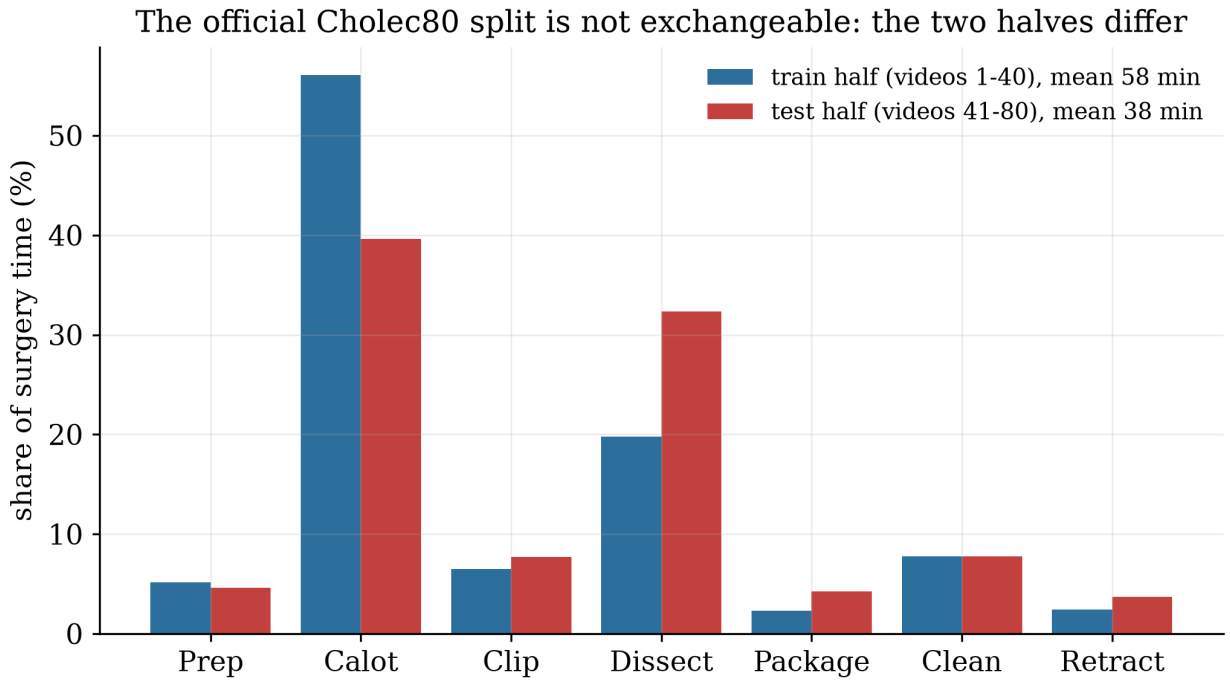

Figure 2: The two halves of the official EndoNet split are not exchangeable. Training surgeries average 3468 seconds and spend 56% of the time in Calot dissection; test surgeries average 2262 seconds with a different phase mix.

with a distribution shift. Both are cheap to run, both are reusable on any phase-labeled surgical dataset, and both change how a headline accuracy number should be read.

Everything reproduces on a laptop, CPU only. The phase labels are non-commercial and are not redistributed; a fetch script plus a checksum manifest rebuild them byte for byte.

Novelty gate

The novelty gate is a kill switch, not prose. A published discovery names the closest prior work and the gap each leaves open; if the question were already answered, the discovery would be killed before any experiment ran.

1. **Twinanda et al., EndoNet / PhaseNet (Cholec80 phase benchmark).** Establishes the benchmark and the vision baselines but never measures a vision-free, time-position-only floor.
2. **Temporal-prior and timestamp-supervision literature for phase recognition.** Adds positional encodings and timestamp supervision to vision models; does not isolate and quantify the standalone time-position shortcut.
3. **Metric-critique work on surgical-phase evaluation.** Critiques the metrics but does not publish a content-free baseline that says how much of a reported number is the schedule.

Reproduce

CPU only. Everything reproduces on a laptop.

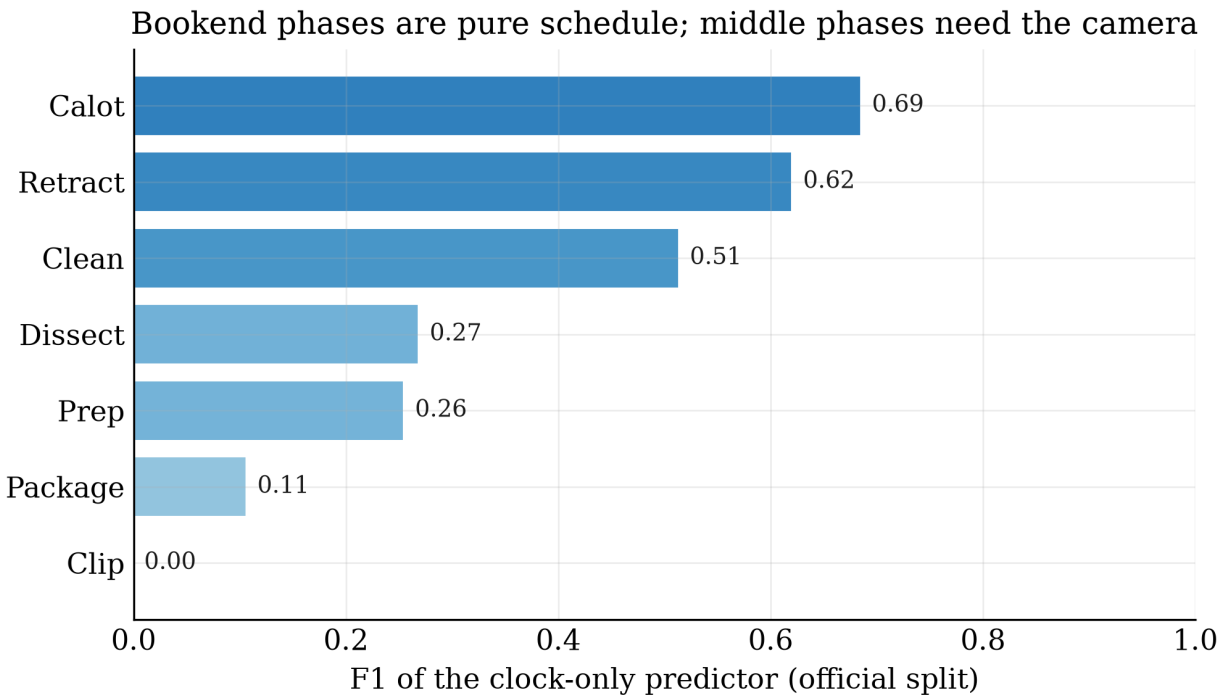

Figure 3: The clock is phase-specific. It nails the long, stereotyped bookend phases (Calot F1 0.69, Retraction 0.62) and is blind to the short interior phases that genuinely need the camera (Clipping F1 0.00).

```
git clone git@github.com:truffle-dev/sd-phase-recognition-temporal-prior.git
cd sd-phase-recognition-temporal-prior
python src/fetch_data.py # reconstruct the 1 fps phase timeline + checksums
python src/content_free_baseline.py # writes results/content_free_metrics.json
python src/make_figures.py # writes the three figures
```

Availability and license

 The canonical record, the machine-readable metadata, and this paper live at https://truffle.help/d/008-cholec80-temporal-prior. **Keywords:** surgical-video, benchmark-validity, shortcut-learning, content-free-baseline.