Reconstructing the missing decks: recovering slide evidence from video-only conference talks

A practice note on the hunt, the framegrab, and the three ways a camera sees a slide

Internet History Initiative

August 2026

The problem

Of the 11,688 video recordings in this collection's talk holdings, 11,240 — ninety-six percent — have no accompanying slide deck. The deck is usually where the substance lives: the numbers, the network diagrams, the timeline the speaker is narrating. A video without its deck is searchable only by its title; the archive knows a talk happened but not what it said.

The curator's proposal, piloted and then batch-tested the same day: first *hunt* for the original deck on the public web; failing that, *reconstruct* the deck from the video itself — framegrab the slides as the camera saw them, and let the collection's page-level apparatus treat the result like any other deck.

Step one: the hunt

The hunt is a per-talk web search (title, plus presenter when the metadata offers a real name) with slide-oriented terms, escalating to event-page fetches and Wayback CDX lookups. Two sessions of results:

When the talk is conference-circuit, the hunt works immediately. The pilot subject — Scott Bradner's "A History of the Internet" (Harvard, 2019) — was found on the first query, archived by the Berkman Klein Center on the Internet Archive. A production find followed the same day: a 53-slide Hurricane Electric peering deck on SlideShare, which also exposed a catalog misspelling of the presenter's name ("Walney" for Wollny) — deck hunts double as metadata audits.

When the source is a streaming channel, the hunt mostly reports honest absence. Forty talks from a streamed-events channel yielded one found deck, twelve confident "no deck exists" verdicts (panels, welcome remarks, an award induction, a radio simulcast), and twenty-seven misses. The verdict taxonomy matters: a *probably-no-deck* with an identified event page is a completed piece of scholarship, not a failure.

Two systematic cautions emerged. Video metadata lies: upload years postdate talks by up to a decade, and one meeting-tagged sample turned out to contain talks from a meeting a decade earlier — every binding must re-verify its event against an authoritative archive, where the deck

often already sits in our own mirrors under the correct meeting. And hunters must be forbidden from constructing URLs: every reported location must have been actually seen.

Step two: the reconstruction

The pipeline is deliberately plain: download the video stream (720p suffices), sample one frame every five seconds, cluster frames into *slide-states* by perceptual hash (16×16 average hash, Hamming distance ≤ 14), keep the sharpest frame of each state, and drop states shorter than ten seconds — those are camera cuts, not slides. Each surviving page records the video timestamp span it was visible, so every reconstructed page carries a working `?t=` link into the moment of the talk where the speaker is saying it. About four minutes of CPU per talk.

The three ways a camera sees a slide

A vision pass over 1,023 reconstructed pages — one model call per page, classifying the frame and transcribing its slide text verbatim — found the population divides three ways:

switched-feed slide (the production cut to the deck)	282
projection-in-scene (the room camera sees the screen)	347
no slide (people, stage, audience)	377
unparsed	17

The projection-in-scene class was the surprise: talks whose metadata said "panel" turned out to contain legible keystoned decks on the room's screen, transcribable as-is. The vision gate also corrected the mechanical verdicts — clustering alone called 36 of 40 talks "reconstructed," but four-plus *slide-bearing* pages is the honest bar, which 29 met; eleven were confirmed slide-free.

The quality inversion

The pilot's sharpest finding: for the Bradner talk we hold both the found deck and the reconstruction, and *the reconstruction is the better artifact*. The archived PDF is a print-resolution two-up handout; the video's switched feed shows each full slide at 720p. Reconstruction is not merely a fallback for missing decks — for switched-feed recordings it can be the archive's best edition, with the found deck serving as the authority on slide count and order.

Provenance discipline

A reconstructed deck is a derived artifact and must never masquerade as the original. Each carries: the source video URL and its SHA-256; the full method string (sampling rate, hash parameters, thresholds); per-page timestamp spans; the vision model that transcribed each page; and a verdict field that distinguishes `reconstructed` from `no-slides-confirmed` and `download-failed`. In the catalog such decks bind as `kind=deck-reconstructed` at receipt tier — machine work, awaiting no one's endorsement, honest about what it is.

Costs and scaling

Per thousand talks: roughly 70 CPU-hours of download-and-cluster (embarrassingly parallel; slurm-array friendly, with downloads kept on the archive's own connection for rate courtesy), and one vision call per surviving page on the collection's standing 35B pool — the model in daily service proved vision-capable, so no dedicated lane was needed. The near-term improvements, in order of value: a perspective-crop rung for the projection-in-scene class; local title-matching against our own mirrors before any web hunt; and event re-verification as a hard gate before binding.

What this buys the reader

A patron who searches for a phrase a speaker only ever put on a slide — never in an abstract, never in a mailing-list post — can now find the talk, land on the reconstructed page, and click through to the second of video where the speaker says it aloud. The deck was always in the video. The archive just had to learn to see it.