
Evaluating Multimodal Narrative Understanding of Popular Hollywood Films

David Bamman,¹ Kent K. Chang,¹ Allison Cooper,² Juishan Hsu,³ Reina Kushihashi,³ Madison Mar,³ Arnav Podichetty,³ Rachael Samberg,⁴ Ipek Nil Sancak³ and Yuhan Shao³

¹School of Information, UC Berkeley ²Cinema Studies, Bowdoin College

³UC Berkeley ⁴Scholarly Communication and Information Policy, UC Berkeley

{dbamman,kentkchang}@berkeley.edu

Abstract

Multimodal language models increasingly show promise for enabling the large-scale computational analysis of film, opening up new avenues for learning about film history and the evolution of narrative techniques. But the creation of stable benchmarks built around Hollywood films is complicated by copyright protections. In this work, we address these concerns directly, by building a new collection of Hollywood films defined by two criteria: box office popularity (where we publish the first large-scale, open collection of weekly box office earnings reported by *Variety* magazine from 1922–1979); and likely public domain status (by researching copyright registrations and renewals in the US *Catalog of Copyright Entries*). We build a new multimodal MCQ benchmark on top of this collection that focuses on narrative elements that directly evaluate the abilities of models to inform meaningful research on film narrative; we find that many vision-language models struggle on this task (with many performing at near-chance levels of accuracy), while audio-visual models (including those that use audio in captioning scenes) reach a maximum accuracy of 61.1%, well below human-level performance.

1 Introduction

One of the most exciting consequences of advances in NLP, CV and AI has been in their application as analytical tools to shed light on questions of culture at scale. While this use has long focused on the domain of text [Underwood, 2019, Piper, 2019], we see increasing application for the analysis of film. This work has shed light on changes in pacing and luminosity [Cutting et al., 2011], the representation of race and gender on screen [Guha et al., 2015, Arnold et al., 2019, Bamman et al., 2024], and comedic timing [Zribi et al., 2026], along with many others [Somandepalli et al., 2021].

However, while progress in other application areas of these methods can be driven by the formation of open benchmarks [Deng et al., 2009, Lin et al., 2014, Kay et al., 2017], film as an object of study presents distinct challenges due to the copyrighted nature of the underlying materials. While many of the core tasks—from shot boundary segmentation [Zabih et al., 1995, Soucek and Lokoc, 2024] to character identification [Everingham et al., 2006, Nagrani et al., 2018]—are often designed with application to the film industry in mind, restrictions on the digitization and republishing of copyrighted materials often leave researchers without direct access to the underlying data. Even if extracting clips for research purposes is a fair use—an exception to copyright owners’ exclusive rights—film companies have a robust licensing market and often challenge uses that should be considered fair. Datasets that instead rely on practitioners downloading videos from URLs on YouTube also find themselves with increasingly more restricted access over time, as videos are removed by users or through compliance with Digital Millennium Copyright Act (“DMCA”) takedown requests. Benchmarks built on this data are unstable.

One solution to this challenge is to build benchmarks on movies in the public domain. The public domain includes not only works currently published before 1931 in the US,¹ but also any film whose copyright owner (typically the production studio) did not renew its copyright during the period in US copyright law when such registration renewals were both available and required. This, however, raises its own challenges: first, there is no known registry of public domain materials, and films that an online source might claim to be in the public domain can be contested; many of the feature films found on sources like the Internet Archive are still in copyright and could be subject to a DMCA takedown request. Second, while thousands of movies are in the public domain, not all of them are equally notable—the public domain includes war propaganda released by the US government, films whose production studios failed to renew their copyright due to lack of interest in the film, and so on. In building a benchmark around Hollywood films, we want to incorporate some measure of cultural significance as well.

In this work we address these critiques through two interventions in data sourcing: first, determining the cultural impression of films in the United States (measured by *popularity* at the US box office); second, identifying films that are truly likely to be in the public domain—not through trusting the claims of a third party, but through researching that status with the US *Catalog of Copyright Entries*. These two criteria—cultural significance and openness—differentiate our work from related efforts to build datasets that include elements of Hollywood films [Tapaswi et al., 2016, Rawal et al., 2024, Wang et al., 2025], including historical ones [Zaranis et al., 2025].

Given the dataset defined by these criteria, we build a benchmark around it for measuring the performance of multimodal models at the task of long-form *narrative* understanding, focusing in particular on questions of temporality, plot, character, setting, perspective, representation, symbolism and object identification. This benchmark lets us assess the performance of a range of models at measurement tasks increasingly driving scholarship in computational social science and the digital humanities. While previous benchmarks have explored long-form motion pictures [Zaranis et al., 2025, Wang et al., 2025] and interpretive tasks in other modalities like text [Sui et al., 2025, Hamilton et al., 2026], we bring these two paradigms together in this work to assess how such models can inform our analytical understanding of complex narrative phenomena in film.

Our work therefore makes the following contributions:

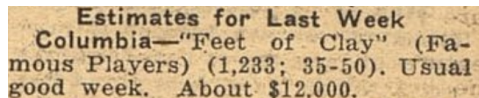
- We present the first systematic database of historical box office earnings (extracted from *Variety* magazine), spanning 1922–1979. This database is publicly available at <https://github.com/bamman-group/variety-boxoffice>.
- We present a new dataset of popular films that are likely in the public domain in the United States, which can form the basis for benchmarks that others can trust will be stable over time.
- We present a new benchmark for narrative understanding in these films, and assess the performance of several multimodal models at this difficult task. We find that many vision-language models struggle on this task (with many performing at near-chance levels of accuracy), while audio-visual models (including those that use audio in captioning scenes) reach a maximum performance of 61.1%, well below human-level performance. This Classical Hollywood Narrative Benchmark (and code to support it) is publicly available at <https://github.com/bamman-group/chnb>.

2 Defining the collection

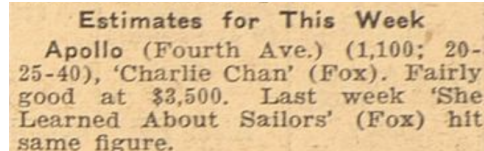
2.1 Identifying popular movies

Our first goal is to identify movies that are popular. Historical box office data in the United States is fragmentary, with major sources of information coming from ledgers kept by executives at individual studios [Glancy, 1992, Jewell, 1994, Glancy, 1995] and trade magazines such as *Variety*, *The Motion Picture Herald* and *Hollywood Reporter*. *Variety* has the deepest historical collection of box office information: starting with its March 3, 1922 issue (and persisting through the early 21st century), *Variety* reports the weekly box office receipts for individual movies in specific theaters. Figure 1 gives two typical examples of this information.

¹Copyright protection extends for a fixed period of time. Films are typically works of corporate authorship protected for 95 years from publication; as of 2026, films published prior to 1931 are in the public domain. The matter is complicated,



Estimates for Last Week
Columbia—“Feet of Clay” (Famous Players) (1,233; 35-50). Usual good week. About \$12,000.



Estimates for This Week
Apollo (Fourth Ave.) (1,100; 20-25-40), “Charlie Chan” (Fox). Fairly good at \$3,500. Last week “She Learned About Sailors” (Fox) hit same figure.

Figure 1: Weekly box office information from two issues of *Variety*, October 29, 1924 (left) and July 10, 1934 (right).

Within a broader column noting the city (Washington and Indianapolis), each entry lists the theater, movie title(s) and box office estimate for that week. We extract this information as a structured tuple—e.g., $\langle \text{Washington, Columbia, Feet of Clay, } \$12,000 \rangle$ —from all pages of *Variety* where the full text of at least one article on the page contains the phrase “estimates for this week”, “estimates for last week”, or where the title of an article contains the phrase “picture grosses”. We prompt multimodal LLMs to extract all tuples from the scan of each page, providing two shots (input image, output json) to illustrate the desired behavior.

Movies mentioned in *Variety* are often terse, relying on readers’ general familiarity with movies that year (e.g., “Gone” to denote *Gone with the Wind* in 1940). For each year, we manually create a mapping of aliases (“Gone with the Wind”, “Gone”, “Gone with Wind”, etc.) to an IMDB record for that movie for all aliases appearing at least 3 times in a year. This process allows us to generate a ranked list of movies each year by their total box office numbers reported by *Variety*.

To evaluate the accuracy of this extraction method, we create a gold-standard dataset by manually labeling 4,072 tuples in 21 issues of *Variety* spanning the years 1922–1979. We use 7 of these issues for development and hold out 12 issues for held-out evaluation. We primarily assess validity through the Spearman rank correlation coefficient between the resulting ranked lists of movies (comparing the ranks derived from human tuples vs. model predicted tuples), since the rank ultimately constitutes the decision criterion we use to define the collection below (the top 100 movies per year). We find that Gemini 3 Pro (with high thinking and ultrahigh image resolution) performs best on this held-out evaluation, with $\rho = 0.961$. The complete evaluation procedure can be found in Appendix B.

We use this best performing model to extract this structured information for all *Variety* issues from 1922–1979. By aggregating the total box office earnings for all movies, we are able to assemble ranked lists of the top grossing movies each year over this complete time frame, comprising 1.4M weekly box office numbers for over 24,000 movies. This represents, to our knowledge, the first comprehensive, openly available dataset of historical box office information for feature films. All extractions, alias mapping, and aggregated yearly/weekly charts are openly available at <https://github.com/bamman-group/variety-boxoffice>.

These top charts allow us to define a measure of popularity, including a film within the boundaries of our dataset if it is among the top 100 highest-grossing films in a year. Commercially distributed features form a logical corpus for a benchmark focusing on narrative understanding because they were produced in industrial conditions that prioritized narrative coherence and broad legibility—the properties we seek to evaluate. The *Variety* dataset is rich in films from the Classical Hollywood era, broadly defined as the period spanning 1917–1960 [Bordwell et al., 1985] and marked by industrial and aesthetic developments such as the studio system and the continuity system, created to improve the clarity and efficiency of storytelling onscreen. Other dominant narrative trends of the period include character-centered causality, goal-oriented plots, and an emphasis on narrative closure [Bordwell, 1985], along with the evolution of the star persona [Dyer, 1998], strategies to navigate the Hollywood Production Code [Jacobs, 1997], and the formation and reinforcement of genre conventions [Altman, 1999]. The dataset spans several years of the Post-Classical period as well; this era saw a loosening of character-centered causality [Elsaesser et al., 1975], increasingly ambiguous endings [Bordwell, 2006], the reworking of Classical-era genres [Ray, 2020], and the growing influence of European art cinema [Bordwell, 1979].

Finally, we note that the limitations of using box office as a proxy for significance are well documented [Maltby, 2006, Staiger, 1992]. Film scholars have cautioned against an over-reliance on commercial

however, by the fact that elements of films may be remastered, and such changes to the original film would be protected by new copyright.

metrics, arguing that too narrow of an approach misses aesthetically important works [Johnston, 1973, Rich, 2013] or the broader historical and social context of cinema [Allen and Gomery, 1985, Verhoeven et al., 2019]. Using box office gross as the main criterion in assembling our collection of popular films means our corpus excludes films that circulated outside mainstream distribution in the United States, from the “race” films of the silent era [Stewart, 2005] to transnational festival films [White, 2015]. It is worth emphasizing that box office is but one of many measures of significance that might be adopted to assemble a film dataset to evaluate multimodal understanding of narrative. Our methodology for identifying the public domain status of popular films, described in detail in the following section, can easily be adapted for the development of benchmarks with different boundaries.

2.2 Investigating public domain status

In the United States, most movies published prior to 1931 have fallen into the public domain (as of the time of this writing); as have most movies whose copyright was registered prior to 1964 but failed to be renewed 28 years later. While the first criterion allows us to identify likely public domain movies relatively easily by considering their year of publication, the latter criterion is more difficult since it requires a.) identifying the year of copyright registration for a film and b.) identifying its *lack* of renewal (including the renewal of any screenplays or expressive works from which the film was adapted, which may bear separate registration).²

We turn to two sources for identifying this information: the *Catalog of Copyright Entries*, published by the US Copyright Office through 1978; and the Copyright Public Records System (CPRS), an online database published by the US Copyright Office, which contains registrations (and renewals) from 1978 forward. To identify candidate movies that may be in the public domain for lack of registration renewal, we extracted all registrations and renewals for movies recorded in the print *Catalog of Copyright Entries* using digitized versions on the Internet Archive (which captures registrations/renewals until 1978) and extracted all motion picture renewals from the Copyright Public Records System to capture renewals made after 1978.

Using this information, we matched movies between the REGISTERED set and the RENEWED set using the original registration number (which often appears as a canonical identifier in both sets); any movie that was matched was automatically excluded, since this provides evidence that the movie was likely appropriately renewed. Since registration numbers may change—whether through OCR mistakes, mistyping, or other factors—we also carried out a detailed manual review of the movies that did not match, attempting to match them based on the similarity of their title.

We also check for one other situation that would lead a movie that has not been renewed to still be under copyright: even if the copyright for the film was not renewed, if the source (such as a short story or novel) was appropriately copyrighted and renewed, then the underlying story for the movie may still remain in copyright as well (e.g., as is the case for *It’s a Wonderful Life*). We draw on data from the American Film Institute, which provides information about whether a movie was based on some other original (e.g., literary) source; any movie described by AFI as being based on an additional source in copyright was removed from the collection.

The two criteria laid out above—popularity (among the top 100 movies per year by box office revenues) and likely public domain status—provide the conceptual boundaries for this collection. We adopt these stringent criteria in order to minimize DMCA takedown requests, and source the content of the films themselves from the Internet Archive, further limiting the collection to only sound films (i.e., no silent-era movies). This results in dataset of 61 popular movies.

3 Building a benchmark

Given this collection of popular films whose copyright status are unlikely to be challenged, we build a benchmark around it to assess the long-form narrative understanding capabilities of multimodal models. We focus in particular on questions that illustrate the affordances of such models for work in cultural analytics of film. For ease of evaluation, we frame the task as a multiple-choice question

²Copyright renewal matters only for a certain time periods. Beginning with the Copyright Act of 1976 (effective January 1, 1978), neither registration nor renewal is required for films to be protected by copyright. Prior to the 1976 Act, however, there is a complex landscape of protection based on a combination of authorship (individual vs. corporate), publication status, registration date, and renewal.



Figure 2: Example question (*His Girl Friday*): “How is the camera positioned as Hildy enters the room to talk to Earl Williams?” A.) On Hildy’s side of the bars to look through the grate at Earl. B.) The camera is at eye level and moves parallel to her. C.) On the inside of the cell to look out at Hildy. D.) High angle above Hildy and Earl.

format, with four answer options (only one of which is correct). We focus on eight narrative categories described below:

- **Temporality.** Questions that track attention to the order of events—both as they occur chronologically within the story world and as they are depicted to the viewer; these orderings are in tension in cases of anachrony [Genette, 1980], such as flashbacks and flashforwards.
Example question: What is the order in which the four characters are arrested? A.) Countess de Mavon → Nurse Edith Cavell → Mme. Moulin → Mme. Rappard. B.) ...
- **Plot.** Questions that identify the narrative function of a scene, whether it introduces a complication, raises the stakes, resolves the central tension, and to track whether character goals stated early in the film are ultimately fulfilled.
Example question: Why didn’t Charles leave the crime scene right away after murdering Meinike? A.) He is setting up a paper trail; B.) ...
- **Character.** Questions that identify roles and track character dynamics based on what the film shows: what characters do, say, how they are dressed, and how they are staged relative to one another.
Example question: Which character is shown smoking? A.) Jack; B.) ...
- **Setting.** Questions that identify and distinguish between locations in the film, and observe how the physical staging of characters within a space (e.g. social blocking) conveys power, relationship, and intention, which require attention to *mise-en-scène* rather than plot.
Example question: Which of these locations do we not see the interior of? A.) King Little’s castle; B.) ...
- **Perspective.** Questions that identify from whose vantage point events are presented, and whether the film ever gives the audience information the characters themselves do not have. This tests attention to how the film is narrated—not what happens, but who knows what, and when.
Example question: When Norma tells Michael how much she loves him, who sees Dr. Besant enter the room first? A.) Norma; B.) We do as viewers (before any characters); C.) ...
- **Representation.** Questions that identify how the film depicts gender roles, social identity, and group membership, based on what is shown and said in the film.
Example question: Does this movie pass the Bechdel test? (Two named women talking to each other about a topic that is not a man.) A.) Yes, between Irma and Phyllis repeatedly ... B.) ...
- **Symbolism.** Questions that identify objects, sounds, or visual actions that recur across the film and carry narrative or thematic weight.
Example question: In Marilyn’s performance where she is surrounded by dishes that she is washing, what does this chore symbolize, based on the lyrics to the song she sings? A.) Growing up; B.) ...
- **Object identification.** Questions that identify a specific on-screen object, track where it appears or what is done with it, or connect it to its function in the plot.
Example question: What object is repeatedly used by neighbors to cope with the heat? A.) Hand fans; B.) ...

To create benchmark questions, seven co-authors viewed the entirety of a movie and created an average of 12.8 questions for each one, resulting in an initial set of 779 questions across 61 films. We

use a plain-language scene description to refer to specific scenes (not explicit timestamps) and avoid distractors that are obviously off-topic to avoid simply testing commonsense reasoning capability. We assess expert-level human performance by distributing questions for a sample of ten movies (133 questions) to co-authors who did not write those questions, asking them to watch the movie and answer all questions (going back and forth to the movie as needed); we find human-level accuracy to reach 82.0%, reflecting in part the complex nature of narrative inferences. Sources of error include ambiguity in the question/answer options (where multiple choices could be argued to be correct), but also reflects the natural difficulty of some information-seeking questions (e.g., where attention is required to a scene that is easily missed). Table 4 (Appendix D) lists the distribution of annotated categories, with greatest representation of questions around plot, character and setting.

4 Memorization

One of the challenges of working with popular movies is that they are frequently discussed online, and these discussions make their way into the pre-training data for LLMs. Past work has found this to be the case as well: Zaranis et al. [2025] report an accuracy of 66.3% (compared to a random performance of 50%) when prompting Gemini 2.5 Pro to answer questions based on the movie title and date of release alone (with no access to the video); Asadi et al. [2026] find this kind of “mirage reasoning” prevalent in multimodal medical benchmarks. We see this as an example of test data contamination [Dodge et al., 2021, Chang et al., 2023], where models use metapragmatic information *about* a movie—rather than the content of the movie itself—to make decisions.

To account for this, we pass all questions through three frontier LLMs—Gemini Pro 3.1, Claude Opus 4.7 and GPT 5.5—with the following prompt: “Based on your knowledge of the movie {MOVIE} ({YEAR}), answer the following question.” All models exhibit similar rates of memorization (Gemini 39.4%, Opus 40.7%, GPT 38.8%). To mitigate this effect, we subselect questions from the pool so that the performance across all models when prompted with the movie title and date alone is approximately 25% (reflecting a random guess), detailed in Appendix D. This yields a total of 628 benchmark questions (discarding 151 from the original pool). As Table 4 (Appendix D) illustrates, the exclusion rate varies by category: models have internal knowledge of common analytical discussion topics about the film—including representation and symbolism—and much less so about questions that require access to specific visuals. For convenience, we let \mathcal{B} denote the post-filter benchmark of 628 questions used in subsequent experiments.

5 Experiments

5.1 Setup

We evaluate five paradigms on \mathcal{B} . For a film with video V (frames and audio) and a question q , each paradigm involves a model M that predicts an answer $\hat{a} = M(c, q)$ from a context c that varies by how V is compressed:

Closed-book baseline: $c = \emptyset$. The system answers from q and its parametric knowledge alone. This establishes whether the benchmark is solvable without the movie at all, a precondition for any subsequent gain to be attributable to movie content rather than priors (visual and textual).

Subtitles-only baseline: $c = \mathcal{S}$. The QA backbone takes only the dialogue transcript \mathcal{S} and the question, without visual or frame-derived input. \mathcal{S} is generated by transcribing the audio track using Distil-Whisper large-v3 [Gandhi et al., 2023]. Prior work on long-movie comprehension has found subtitle-only access to be a surprisingly strong baseline [Zaranis et al., 2025]. We include it as a baseline both for performance comparison and to isolate dialogue as a separate point for long-video compression strategies for narrative understanding.

End-to-end: $c \subseteq V$. The QA model is itself a multimodal model and extracts a fixed sample of V directly. We consider the following types of end-to-end models: a.) *Long-video models* are architectures designed for long-form video and extracts a uniform 64-frame sample of the film. b.) *Vision-language models* are general-purpose VLMs and take a uniform 256-frame sample. c.) *The audio-visual model* (Gemini 3 Flash) processes the video in its entirety, including the audio track.

Socratic [Zeng et al., 2023]: $c = \text{CAPTIONER}(V)$. We follow the two-stage pipeline described in Chandrasegaran et al. [2024]: First, a CAPTIONER produces one of the following: a.) *Frame captions* are generated from a 0.5-fps frame strip with no audio; the captioner sees roughly thirty stills per minute of film. b.) *Clip captions* are generated from video chunks (target duration 60 seconds) that respect shot boundaries [Soucek and Lokoc, 2024] and include the audio track. These captions are then timestamped and concatenated chronologically into a world state history [Chandrasegaran et al., 2024] that the QA backbone receives (instead of V). This class tests whether textual compression preserves the audio-visual signal for long-video QA. We adopt Gemini 3 Flash as the captioner.

Agentic retrieval: $c = \pi(V, q)$. To test whether query-conditioned selectivity improves on the Socratic baseline (which has access to full content), we instantiate the retrieval policy π in two ways: a.) *Frame image retrieval* is the VideoAgent loop described in Fan et al. [2025]: π starts from 8 uniform frames in a 64-frame pool and re-fetches up to four more per iteration via CLIP [Radford et al., 2021] cosine similarity on a confidence-threshold loop (≤ 3 iterations). b.) *Caption retrieval* utilizes the Letta agent [Packer et al., 2023], where π loads the Socratic captions (both frame- and clip-based) into archival memory and lets the QA backbone decide to trigger a semantic search.³

Pre-processing pipeline. Several film-level artifacts are precomputed once per film and shared across paradigms: a.) Uniform frame samples at $N \in \{64, 128, 256\}$ are extracted at indices $\lfloor i \cdot T/N \rfloor$ where T is the film’s frame count, and cached as JPEGs. b.) Shot-grouped chunks come from detected shot boundaries merged greedily to a target duration of 60 seconds, with a 10-second minimum and the constraint that no shot is split. c.) Per-frame CLIP embeddings of the largest uniform pool, L_2 -normalized and cached, drive the targeted frame-retrieval step in the VideoAgent loop. d.) Captions are produced by Gemini 3 Flash from either the 0.5-fps frame strip (frame captions) or the shot-grouped video chunk with audio (clip captions).

5.2 Results

We evaluate \mathcal{B} on the following: HourLLaVA [Lin et al., 2025], VAMBA-Qwen2-VL-7B [Ren et al., 2025], VideoChat-Flash [Li et al., 2026], and LLaVA-NeXT-Video-DPO [Zhang et al., 2024] are long-video models; Qwen3-VL-8B [Bai et al., 2025] and GLM-4.1V-9B-Thinking [Hong et al., 2026] are open-weight VLMs; and finally, closed-source models: GPT-5-mini [Singh et al., 2025], Claude Haiku 4.5 [Anthropic, 2025], and Gemini 3 Flash [Gemini Team et al., 2023].⁴ Table 1 summarizes model performance. For each setup, we report accuracy and note the widest 95% Wald confidence intervals to facilitate testing the significance of direct model comparisons.

In the closed-book baseline, no lower CI bound exceeds 25%; models cannot answer questions in \mathcal{B} from their parametric knowledge alone. Nor do the closed-source backbones agree on which films they know better: per-film closed-book accuracies correlate at $\rho = +0.11$ ($p = 0.40$) between Gemini Flash 3 and GPT-5-mini, $+0.28$ ($p = 0.027$) between Gemini Flash 3 and Claude Haiku 4.5, and $+0.44$ ($p < 0.001$) between GPT-5-mini and Claude. The subtitle-only baseline shows that the questions are meaningfully answerable from real movie content; dialogue alone (the ASR transcript) raises accuracy to 48.1 for Gemini, 39.2 for GPT, and 36.9 for Claude.

Clip-based Socratic captioning is competitive with native long-video processing. The strongest end-to-end configuration is Gemini 3 Flash on full video (61.1 ± 3.8), followed by the same model on clip-based captions (58.9 ± 3.8). Given the overlapping CIs, there is no meaningful gap between the text-mediated compression via clip captioning and Gemini’s native multimodal processing. Clip-based Socratic captioning is effective across models, which makes an empirical case for caption-based compression as a viable alternative to video processing.

Performance of agentic and Socratic methods differs across QA backbones. The Letta agent and Socratic pipelines take the same captions but differ in how they reach the model: Letta retrieves from archival memory, but Socratic includes the entire chronological description based on the captions in the prompt. The Letta–Socratic gap is significant for Gemini on both caption types and for GPT on

³<https://docs.letta.com/api/python/resources/agents/subresources/passages/methods/search/>. We use the default `text-embedding-3-small` model for semantic search.

⁴Gemini 3 models are accessed via Vertex AI: <https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models>.

Table 1: Accuracy on \mathcal{B} . The largest 95% Wald confidence interval is $\pm 3.9\%$. Bold indicates best overall; underline marks the within-subgroup leader significantly better than the runner-up (whose 95% CIs do not overlap).

Model	Compute	Acc.	Model	Compute	Acc.
<i>Closed-book</i> (question only)			<i>Socratic</i>		
Qwen3-VL-8B	≈0h 07m	21.5	FRAME-BASED		
GLM-4.1V-9B-Thinking	≈2h 46m	20.4	Qwen3-VL-8B	≈1h 36m	30.3
GPT-5-mini	≈0h 03m	25.2	GPT-5-mini	≈4h 42m	36.8
Claude Haiku 4.5	≈0h 09m	22.1	Claude Haiku 4.5	≈0h 58m	35.0
Gemini 3 Flash	≈2h 47m	26.6	Gemini 3 Flash	≈7h 15m	<u>46.0</u>
<i>Subtitles only</i>			CLIP-BASED		
Qwen3-VL-8B	≈0h 20m	32.2	Qwen3-VL-8B	≈1h 44m	32.8
GLM-4.1V-9B-Thinking	≈31h 21m	27.7	GPT-5-mini	≈3h 30m	48.1
GPT-5-mini	≈9h 22m	39.2	Claude Haiku 4.5	≈1h 02m	45.9
Claude Haiku 4.5	≈0h 08m	36.9	Gemini 3 Flash	≈2h 53m	<u>58.9</u>
Gemini 3 Flash	≈2h 04m	<u>48.1</u>	<i>Agentic retrieval</i>		
<i>End-to-end</i>			FRAME IMAGE (VideoAgent)		
LONG-VIDEO (64 frames)			Qwen3-VL-8B	≈16h 00m	22.9
VAMBA-Qwen2-VL-7B	≈1h 47m	23.7	GPT-5-mini	≈9h 22m	24.8
VideoChat-Flash	≈1h 04m	27.2	Claude Haiku 4.5	≈1h 53m	26.1
HourLLaVA	≈1h 02m	23.4	Gemini 3 Flash	≈2h 04m	28.8
LLaVA-NeXT-Video-7B-DPO	≈0h 15m	25.0	FRAME CAPTION (Letta)		
VISION-LANGUAGE (256 frames)			Qwen3-VL-8B	≈1h 31m	26.1
Qwen3-VL-8B	≈4h 43m	29.1	GPT-5-mini	≈1h 57m	33.6
GLM-4.1V-9B-Thinking	≈6h 54m	24.2	Claude Haiku 4.5	≈3h 27m	<u>39.8</u>
GPT-5-mini	≈7h 52m	34.2	Gemini 3 Flash	≈4h 48m	26.6
Gemini 3 Flash	≈6h 22m	39.3	CLIP CAPTION (Letta)		
AUDIO-VISUAL (full video)			Qwen3-VL-8B	≈1h 50m	26.8
Gemini 3 Flash	≈37h 24m	61.1	GPT-5-mini	≈2h 13m	26.0
			Claude Haiku 4.5	≈1h 51m	<u>41.2</u>
			Gemini 3 Flash	≈6h 51m	30.3

clip captions; for Claude the two paradigms are tied, and it is the strongest agentic-retrieval backbone of the three. In contrast, frame-image retrieval barely exceeds chance on every backbone. The agent’s iterative frame-retrieval loop cannot compensate for what its modality omits (audio and dialogue). Its CLIP-similarity is unsuitable for retrieving content in the audio stream that is relevant to the answer.

Frame budget exerts limited impact. For the six end-to-end backbones run at multiple frame budgets, we report average accuracy in Table 5 (Appendix E) for $N \in \{64, 128, 256\}$. The largest within-backbone gain is +3.6 pp from $N=64$ to $N=256$; within-backbone CIs overlap heavily across the three budgets, and we therefore use a single canonical budget per sub-paradigm in Table 1 ($N=64$ for long-video models, $N=256$ for vision-language models).

Vision is not sufficient for film narrative understanding. Of the four architectures designed for hour-scale video at 64 frames (HourLLaVA, VAMBA-Qwen2-VL-7B, and VideoChat-Flash, LLaVA-NeXT-Video-DPO), none are statistically indistinguishable from the closed-book baseline. The general-purpose vision-language models with a larger frame budget do better, but in the case of Gemini 3 Flash, switching from a 256-frame visual input (39.3%) to full video, including audio track (61.1%), gains +21.8 pp. We observe similar patterns inside Socratic: holding the QA backbone fixed, swapping the captions from frame-based (no audio) to clip-based introduces significant gains for all three models.

Performance gains come from dialogue access and reasoning over video content. Per-film accuracy in each backbone’s best configuration correlates positively with that backbone’s subtitles-only accuracy (Table 2; $\rho_{\text{subtitles}}$ ranges from +0.40 to +0.52, all $p \leq 0.002$): the films where the strongest configurations win are in the films where dialogue alone is informative. To assess whether

Table 2: Spearman rank correlations across all 61 films, of per-film accuracy in each backbone’s best-performing configuration, compared with four predictors: closed-book accuracy of Gemini 3 Flash, subtitles-only accuracy, web prevalence, and title prediction accuracy.

Model	Best	Acc.	Gemini closed-book		Subtitles		Web hits		Title prediction	
			ρ_{gcb}	P	$\rho_{\text{subtitles}}$	P	ρ_{hits}	P	ρ_{title}	P
GPT-5-mini	Socratic, clip captions	48.1	+0.22	0.084	+0.40	0.002	-0.25	0.06	-0.13	0.31
Claude Haiku 4.5	Socratic, clip captions	45.9	+0.37	0.003	+0.52	<0.001	-0.45	<0.001	-0.16	0.23
Gemini 3 Flash	end-to-end, full video	61.1	+0.43	<0.001	+0.42	<0.001	-0.27	0.04	-0.26	0.05

memorization drives the performance gains on \mathcal{B} , we measure two proxies: web prevalence and title-prediction accuracy. Web prevalence is the $\log_{10}(1 + h_v)$ -scaled count of Google Search results (h_v) for the title query, and title-prediction accuracy captures the fraction of 10 uniformly-spaced 5-frame window per film from which the backbone correctly predicts the canonical film title on IMDb, assessed by case-insensitive exact string match (28.4% for Gemini 3 Flash, 3.3% for GPT-5-mini, and 2.1% for Claude Haiku 4.5). A model whose performance partly relies on memorized content would perform better on films with greater web presence and whose visual iconography it can identify. However, in Table 2, we see that across all three closed models, ρ_{hits} and ρ_{title} are negative, suggesting films that are popular online and visually recognizable by the model are not easier even on the strongest configurations.

Gemini’s $\rho_{\text{gcb}} = +0.43$ is a within-model consistency effect, but for Claude and GPT-5-mini, their clip-captioned Socratic routes the audio-visual content through Gemini Flash as a captioner before reaching the QA backbone. The +0.37 and +0.22 we observe for those two against the closed-book ranking of Gemini Flash, then, show that Gemini’s parametric film knowledge influences its caption output enough to leave a per-film signal in the performance of the downstream models.

6 Conclusion

We present in this work a new benchmark of narrative questions built around popular movies from the Classical Hollywood era, defining the boundaries of that collection by movies that are popular (as measured by box office numbers reported by *Variety* magazine) and whose copyright status is unlikely to be challenged (either by being released prior to 1931 or by registering their copyright but failing to renew it). The questions require attention to complex narrative elements involving plot, character, setting, perspective, and more, and prove challenging for frontier multimodal language models. As more research leverages such models for the large-scale computational analysis of film, we expect this benchmark to provide a proving ground for assessing comparative model performance. Data and code to support this work are available at <https://github.com/bamman-group/variety-boxoffice> and <https://github.com/bamman-group/chnb>.

7 Limitations

While this work aims to address a gap in long-form multimodal benchmarks for assessing narrative understanding abilities of contemporary models, it is limited in several ways. By selecting movies based on popularity alone, we encode only one of the many possible forms of cultural significance, and omit movies from the time period that circulated outside of major metropolitan cities in the United States. The process we describe for investigating public domain status, however, could be applied to define new collections under alternative criteria. Additionally, all questions in the benchmark were created by researchers (of varying disciplinary backgrounds) at U.S. universities, which influences the narrative aspects in a film we find salient. Finally, the benchmark specifically covers the era of Classical Hollywood cinema (through 1963); while we expect models with good long-form narrative understanding to be able to perform well on this data, performance may not generalize to films outside of this time period (both older and newer); we see this as a necessary trade-off for defining a collection of movies on which additional stable benchmarks can be built.

Acknowledgments

The research reported in this article was supported by the Humanities and AI Virtual Institute (HAVI), a program of Schmidt Sciences, and by Google. This research used the Savio computational cluster resource provided by the Berkeley Research Computing program at the University of California, Berkeley.

References

- Robert Clyde Allen and Douglas Gomery. *Film history: Theory and practice*. Knopf, 1985.
- Rick Altman. *Film/Genre*. BFI Publishing, 1999.
- Anthropic. Claude Haiku 4.5 system card. Technical report, Anthropic, October 2025.
- Taylor Arnold, Lauren Tilton, and Annie Berke. Visual style in two network era sitcoms. *Journal of Cultural Analytics*, 4(2), 2019.
- Mohammad Asadi, Jack W O’Sullivan, Fang Cao, Tahoura Nedae, Kamyar Fardi, Fei-Fei Li, Ehsan Adeli, and Euan Ashley. Mirage the illusion of visual understanding. *arXiv preprint arXiv:2603.21687*, 2026.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibao Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.
- David Bamman, Rachael Samberg, Richard Jean So, and Naitian Zhou. Measuring diversity in Hollywood through the large-scale computational analysis of film. *Proceedings of the National Academy of Sciences*, 121(46):e2409770121, 2024. doi: 10.1073/pnas.2409770121. URL <https://www.pnas.org/doi/abs/10.1073/pnas.2409770121>.
- David Bordwell. The art cinema as a mode of film practice. *Film Criticism*, 4(1):56–64, 1979.
- David Bordwell. *Narration in the Fiction Film*. Univ of Wisconsin Press, 1985.
- David Bordwell. *The way Hollywood tells it: Story and style in modern movies*. Univ of California Press, 2006.
- David Bordwell, Janet Staiger, and Kristin Thompson. *The classical Hollywood cinema: Film style & mode of production to 1960*. Columbia University Press, 1985.
- Keshigeyan Chandrasegaran, Agrim Gupta, Lea M. Hadzic, Taran Kota, Jimming He, Cristobal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. HourVideo: 1-hour video-language understanding. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 53168–53197. Curran Associates, Inc., 2024. doi: 10.52202/079017-1684. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/5f2809607f692d79a01c05c43d702883-Paper-Datasets_and_Benchmarks_Track.pdf.
- Kent K. Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. Speak, memory: An archaeology of books known to ChatGPT/GPT-4. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7312–7327, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.453. URL <https://aclanthology.org/2023.emnlp-main.453/>.

- James E Cutting, Kaitlin L Brunick, Jordan E DeLong, Catalina Iricinschi, and Ayse Candan. Quicker, faster, darker: Changes in Hollywood film over 75 years. *i-Perception*, 2(6):569–576, 2011.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 1286–1305, 2021.
- Richard Dyer. *Stars*. 1979. London: BFI, 1998.
- Thomas Elsaesser et al. The pathos of failure: the unmotivated hero. *Monogram*, 6, 1975.
- Mark Everingham, Josef Sivic, and Andrew Zisserman. “Hello! my name is... Buffy” – Automatic naming of characters in TV video. In *BMVC*, volume 2, page 6, 2006.
- Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, pages 75–92. Springer, 2025.
- Sanchit Gandhi, Patrick von Platen, and Alexander M. Rush. Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling, 2023.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, Yaguang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Gra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, Hyunjeong Choe, goston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Meyer, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anas White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maroon, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iaki Iturrate, Ruiho Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adri Puigdomenech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sbastien M R Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodgkinson, Pranav Shyam, Johan Ferret, Steven Hand,

Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo-Yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjöstrand, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlias, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton

Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrz, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaille, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, Z J Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi Lv, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Reynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tume, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian Lin, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnappalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon,

Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, Mohammadhossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser Tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kaffe, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Pöder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A Choquette-Choo, Yunjie Li, T J Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Riviére, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fildjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshv, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldrige, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, Xianghai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurusurthy, Mark Goldenson, Parashar Shah, M K Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirsenschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhanian, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani,

- Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A family of highly capable multimodal models. *arXiv [cs.CL]*, December 2023.
- G rard Genette. *Narrative discourse: An essay in method*, volume 3. Cornell University Press, 1980.
- H Mark Glancy. MGM film grosses, 1924–1948: The Eddie Mannix ledger. *Historical Journal of Film, Radio and Television*, 12(2):127–144, 1992.
- H Mark Glancy. Warner Bros film grosses, 1921–51: The William Schaefer ledger. *Historical Journal of Film, Radio and Television*, 15(1):55–73, 1995.
- Tanaya Guha, Che-Wei Huang, Naveen Kumar, Yan Zhu, and Shrikanth S Narayanan. Gender representation in cinematic content: A multimodal approach. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 31–34, 2015.
- Sil Hamilton, Matthew Wilkens, and Andrew Piper. NarraBench: A comprehensive framework for narrative benchmarking. In Vera Demberg, Kentaro Inui, and Llu s Marquez, editors, *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3786–3801, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.176. URL <https://aclanthology.org/2026.eacl-long.176/>.
- Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihang Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Haochen Li, Jiale Zhu, Jiali Chen, Jiaying Xu, Jiazheng Xu, Jing Chen, Jinghao Lin, Jinhao Chen, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Ruiliang Lyu, Shangqin Tu, Sheng Yang, Shengbiao Meng, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wei Jia, Wenkai Li, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyu Zhang, Xinyue Fan, Xuancheng Huang, Yadong Xue, Yanfeng Wang, Yanling Wang, Yanzi Wang, Yifan An, Yifan Du, Yiheng Huang, Yilin Niu, Yiming Shi, Yu Wang, Yuan Wang, Yuanchang Yue, Yuchen Li, Yusen Liu, Yutao Zhang, Yuting Wang, Yuxuan Zhang, Zhao Xue, Zhengxiao Du, Zhenyu Hou, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. GLM-4.5V and GLM-4.1V-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2026.
- Lea Jacobs. *The wages of sin: Censorship and the fallen woman film, 1928-1942*. Univ of California Press, 1997.
- Richard B Jewell. RKO film grosses, 1929–1951: The CJ Tevlin ledger. *Historical Journal of Film, Radio and Television*, 14(1):37–49, 1994.
- Claire Johnston. *Women’s cinema as counter-cinema*. SEFT, London, 1973.
- Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The Kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.
- Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhang Zhu, Haiyan Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. VideoChat-Flash: Hierarchical compression for long-context video modeling. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=MUjdNcfNPv>.

- Jingyang Lin, Jialian Wu, Ximeng Sun, Ze Wang, Jiang Liu, Yusheng Su, Xiaodong Yu, Hao Chen, Jiebo Luo, Zicheng Liu, and Emad Barsoum. Unleashing hour-scale video training for long video-language understanding. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, pages 17523–17552. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/19741c617c87a25627682e5714af1501-Paper-Conference.pdf.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
- Richard Maltby. On the prospect of writing cinema history from below. *TMG Journal for Media History*, 9(2), 2006.
- Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Learnable pins: Cross-modal embeddings for person identity. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.
- Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G Patil, Ion Stoica, and Joseph E Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv [cs.AI]*, October 2023.
- Andrew Piper. *Enumerations: Data and literary study*. University of Chicago Press, 2019.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. URL <https://proceedings.mlr.press/v139/radford21a.html>.
- Ruchit Rawal, Khalid Saifullah, Miquel Farré, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. Cinepile: A long video question answering dataset and benchmark. *arXiv preprint arXiv:2405.08813*, 2024.
- Robert B Ray. *A certain tendency of the Hollywood cinema, 1930-1980*. Princeton University Press, 2020.
- Weiming Ren, Wentao Ma, Huan Yang, Cong Wei, Ge Zhang, and Wenhui Chen. Vamba: Understanding hour-long videos with hybrid mamba-transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21197–21208, October 2025.
- B Ruby Rich. *New queer cinema: The director's cut*. Duke University Press, 2013.
- Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, A J Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, Alex Makelov, Alex Neitz, Alex Wei, Alexandra Barr, Alexandre Kirchmeyer, Alexey Ivanov, Alexi Christakis, Alistair Gillespie, Allison Tam, Ally Bennett, Alvin Wan, Alyssa Huang, Amy Mcdonald Sandjideh, Amy Yang, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrei Gheorghe, Andres Garcia Garcia, Andrew Braunstein, Andrew Liu, Andrew Schmidt, Andrey Mereskin, Andrey Mishchenko, Andy Applebaum, Andy Rogerson, Ann Rajan, Annie Wei, Anoop Kotha, Anubha Srivastava, Anushree Agrawal, Arun Vijayvergiya, Ashley Tyra, Ashvin Nair, Avi Nayak, Ben Eggers, Bessie Ji, Beth Hoover, Bill Chen, Blair Chen, Boaz Barak, Borys Minaiev, Botao Hao, Bowen Baker, Brad Lightcap, Brandon McKinzie, Brandon Wang, Brendan Quinn, Brian Fioca, Brian Hsu, Brian Yang, Brian Yu, Brian Zhang, Brittany Brenner, Callie Riggins Zetino, Cameron Raymond, Camillo Lugaresi, Carolina Paz, Cary Hudson, Cedric Whitney, Chak Li, Charles Chen, Charlotte Cole, Chelsea Voss, Chen Ding, Chen Shen, Chengdu Huang, Chris Colby, Chris Hallacy, Chris Koch, Chris Lu, Christina Kaplan, Christina Kim, C J Minott-Henriques, Cliff Frey, Cody Yu, Coley Czarnecki, Colin Reid, Colin Wei, Cory Decareaux, Cristina Scheau, Cyril Zhang, Cyrus Forbes, Da Tang, Dakota Goldberg, Dan Roberts, Dana Palmie, Daniel Kappler, Daniel Levine, Daniel Wright, Dave Leo, David Lin,

David Robinson, Declan Grabb, Derek Chen, Derek Lim, Derek Salama, Dibya Bhattacharjee, Dimitris Tsipras, Dinghua Li, Dingli Yu, D J Strouse, Drew Williams, Dylan Hunn, Ed Bayes, Edwin Arbus, Ekin Akyurek, Elaine Ya Le, Elana Widmann, Eli Yani, Elizabeth Proehl, Enis Sert, Enoch Cheung, Eri Schwartz, Eric Han, Eric Jiang, Eric Mitchell, Eric Sigler, Eric Wallace, Erik Ritter, Erin Kavanaugh, Evan Mays, Evgenii Nikishin, Fangyuan Li, Felipe Petroski Such, Filipe de Avila Belbute Peres, Filippo Raso, Florent Bekerman, Foivos Tsimpourlas, Fotis Chantzis, Francis Song, Francis Zhang, Gaby Raila, Garrett McGrath, Gary Briggs, Gary Yang, Giambattista Parascandolo, Gildas Chabot, Grace Kim, Grace Zhao, Gregory Valiant, Guillaume Leclerc, Hadi Salman, Hanson Wang, Hao Sheng, Haoming Jiang, Haoyu Wang, Haozhun Jin, Harshit Sikchi, Heather Schmidt, Henry Aspegren, Honglin Chen, Huida Qiu, Hunter Lightman, Ian Covert, Ian Kivlichan, Ian Silber, Ian Sohl, Ibrahim Hammoud, Ignasi Clavera, Ikai Lan, Ilge Akkaya, Ilya Kostrikov, Irina Kofman, Isak Etinger, Ishaan Singal, Jackie Hehir, Jacob Huh, Jacqueline Pan, Jake Wilczynski, Jakub Pachocki, James Lee, James Quinn, Jamie Kiros, Janvi Kalra, Jasmyn Samaroo, Jason Wang, Jason Wolfe, Jay Chen, Jay Wang, Jean Harb, Jeffrey Han, Jeffrey Wang, Jennifer Zhao, Jeremy Chen, Jerene Yang, Jerry Tworek, Jesse Chand, Jessica Landon, Jessica Liang, Ji Lin, Jiancheng Liu, Jianfeng Wang, Jie Tang, Jihan Yin, Joanne Jang, Joel Morris, Joey Flynn, Johannes Ferstad, Johannes Heidecke, John Fishbein, John Hallman, Jonah Grant, Jonathan Chien, Jonathan Gordon, Jongsoo Park, Jordan Liss, Jos Kraaijeveld, Joseph Guay, Joseph Mo, Josh Lawson, Josh McGrath, Joshua Vendrow, Joy Jiao, Julian Lee, Julie Steele, Julie Wang, Junhua Mao, Kai Chen, Kai Hayashi, Kai Xiao, Kamyar Salahi, Kan Wu, Karan Sekhri, Karan Sharma, Karan Singhal, Karen Li, Kenny Nguyen, Keren Gu-Lemberg, Kevin King, Kevin Liu, Kevin Stone, Kevin Yu, Kristen Ying, Kristian Georgiev, Kristie Lim, Kushal Tirumala, Kyle Miller, Lama Ahmad, Larry Ly, Laura Clare, Laurance Fauconnet, Lauren Itow, Lauren Yang, Laurentia Romaniuk, Leah Anise, Lee Byron, Leher Pathak, Leon Maksin, Leyan Lo, Leyton Ho, Li Jing, Liang Wu, Liang Xiong, Lien Mamitsuka, Lin Yang, Lindsay McCallum, Lindsey Held, Liz Bourgeois, Logan Engstrom, Lorenz Kuhn, Louis Feuvrier, Lu Zhang, Lucas Switzer, Lukas Kondraciuk, Lukasz Kaiser, Manas Joglekar, Mandeep Singh, Mandip Shah, Manuka Stratta, Marcus Williams, Mark Chen, Mark Sun, Marselus Cayton, Martin Li, Marvin Zhang, Marwan Aljubeih, Matt Nichols, Matthew Haines, Max Schwarzer, Mayank Gupta, Meghan Shah, Melody Huang, Meng Dong, Mengqing Wang, Mia Glaese, Micah Carroll, Michael Lampe, Michael Malek, Michael Sharman, Michael Zhang, Michele Wang, Michelle Pokrass, Mihai Florian, Mikhail Pavlov, Miles Wang, Ming Chen, Mingxuan Wang, Minnia Feng, Mo Bavarian, Molly Lin, Moose Abdool, Mostafa Rohaninejad, Nacho Soto, Natalie Staudacher, Natan LaFontaine, Nathan Marwell, Nelson Liu, Nick Preston, Nick Turley, Nicklas Ansmann, Nicole Blades, Nikil Pancha, Nikita Mikhaylin, Niko Felix, Nikunj Handa, Nishant Rai, Nitish Keskar, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Oona Gleeson, Pamela Mishkin, Patryk Lesiewicz, Paul Baltescu, Pavel Belov, Peter Zhokhov, Philip Pronin, Phillip Guo, Phoebe Thacker, Qi Liu, Qiming Yuan, Qinghua Liu, Rachel Dias, Rachel Puckett, Rahul Arora, Ravi Teja Mullapudi, Raz Gaon, Reah Miyara, Rennie Song, Rishabh Aggarwal, R J Marsan, Robel Yemiru, Robert Xiong, Rohan Kshirsagar, Rohan Nuttall, Roman Tsiupa, Ronen Eldan, Rose Wang, Roshan James, Roy Ziv, Rui Shu, Ruslan Nigmatullin, Saachi Jain, Saam Talaie, Sam Altman, Sam Arnesen, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Sarah Yoo, Savannah Heon, Scott Ethersmith, Sean Grove, Sean Taylor, Sebastien Bubeck, Sever Banesiu, Shaokyi Amdo, Shengjia Zhao, Sherwin Wu, Shibani Santurkar, Shiyu Zhao, Shraman Ray Chaudhuri, Shreyas Krishnaswamy, Shuaiqi, Xia, Shuyang Cheng, Shyamal Anadkat, Simón Posada Fishman, Simon Tobin, Siyuan Fu, Somay Jain, Song Mei, Sonya Egoian, Spencer Kim, Spug Golden, S Q Mah, Steph Lin, Stephen Imm, Steve Sharpe, Steve Yadlowsky, Sulman Choudhry, Sungwon Eum, Suvansh Sanjeev, Tabarak Khan, Tal Stramer, Tao Wang, Tao Xin, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Degry, Thomas Shadwell, Tianfu Fu, Tianshi Gao, Timur Garipov, Tina Sriskandarajah, Toki Sherbakov, Tomer Kaftan, Tomo Hiratsuka, Tongzhou Wang, Tony Song, Tony Zhao, Troy Peterson, Val Kharitonov, Victoria Chernova, Vineet Kosaraju, Vishal Kuo, Vitchyr Pong, Vivek Verma, Vlad Petrov, Wanning Jiang, Weixing Zhang, Wenda Zhou, Wenlei Xie, Wenting Zhan, Wes McCabe, Will DePue, Will Ellsworth, Wulfie Bain, Wyatt Thompson, Xiangning Chen, Xiangyu Qi, Xin Xiang, Xinwei Shi, Yann Dubois, Yaodong Yu, Yara Khakbaz, Yifan Wu, Yilei Qian, Yin Tat Lee, Yinbo Chen, Yizhen Zhang, Yizhong Xiong, Yonglong Tian, Young Cha, Yu Bai, Yu Yang, Yuan Yuan, Yuanzhi Li, Yufeng Zhang, Yuguang Yang, Yujia Jin, Yun Jiang, Yunyun Wang, Yushi Wang, Yutian Liu, Zach Stubenvoll, Zehao Dou, Zheng Wu, and Zhigang Wang. OpenAI GPT-5 system card. *arXiv [cs.CL]*, December 2025.

- Krishna Somandepalli, Tanaya Guha, Victor R. Martinez, Naveen Kumar, Hartwig Adam, and Shrikanth Narayanan. Computational media intelligence: Human-centered machine analysis of media. *Proceedings of the IEEE*, 109(5):891–910, 2021. doi: 10.1109/JPROC.2020.3047978.
- Tomás Soucek and Jakub Lokoc. Transnet v2: An effective deep network architecture for fast shot transition detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11218–11221, 2024.
- Janet Staiger. *Interpreting films: Studies in the historical reception of American cinema*. Princeton University Press, 1992.
- Jacqueline Najuma Stewart. *Migrating to the movies: Cinema and Black urban modernity*. Univ of California Press, 2005.
- Peiqi Sui, Juan Diego Rodriguez, Philippe Laban, J. Dean Murphy, Joseph P. Dexter, Richard Jean So, Samuel Baker, and Pramit Chaudhuri. KRISTEVA: Close reading as a novel task for benchmarking interpretive reasoning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32829–32849, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1577. URL <https://aclanthology.org/2025.acl-long.1577/>.
- Makarand Tapaswi, Yukun Zhu, Rainer Stiefelwagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. MovieQA: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016.
- Ted Underwood. *Distant horizons: digital evidence and literary change*. University of Chicago Press, 2019.
- D Verhoeven, B Coate, and V Zemaityte. Re-distributing gender in the global film industry: Beyond #MeToo and #MeThree. *Media Industries*, 2019.
- Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22958–22967, 2025.
- Patricia White. *Women’s cinema, world cinema: projecting contemporary feminisms*. Duke University Press, 2015.
- Ramin Zabih, Justin Miller, and Kevin Mai. A feature-based algorithm for detecting and classifying scene breaks. In *Proceedings of the third ACM international conference on Multimedia*, pages 189–200, 1995.
- Emmanouil Zaranis, António Farinhas, Saul Santos, Beatriz Canaverde, Miguel Moura Ramos, Aditya K Surikuchi, André Viveiros, Baohao Liao, Elena Bueno-Benito, Nithin Sivakumaran, Pavlo Vasylenko, Shoubin Yu, Sonal Sannigrahi, Wafaa Mohammed, Ben Peters, Danae Sánchez Villegas, Elias Stengel-Eskin, Giuseppe Attanasio, Jaehong Yoon, Stella Frank, Alessandro Suglia, Chrysoula Zerva, Desmond Elliott, Mariella Dimiccoli, Mohit Bansal, Oswald Lanz, Raffaella Bernardi, Raquel Fernández, Sandro Pezzelle, Vlad Niculae, and André F. T. Martins. Movie Facts and Fibs (MF²): A benchmark for long movie understanding, 2025. URL <https://arxiv.org/abs/2506.06275>.
- Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Marcin Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael S Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. Socratic models: Composing zero-shot multimodal reasoning with language. In *The Eleventh International Conference on Learning Representations*, 2023.
- Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. LLaVA-NeXT: A strong zero-shot video understanding model, April 2024. URL <https://llava-vl.github.io/blog/2024-04-30-llava-next-video/>.
- Yaelle Zribi, Florian Cafiero, Vincent Lépinay, and Chahan Vidal-Gorène. Timing in stand-up comedy: Text, audio, laughter, kinesics (TIC-TALK): Pipeline and database for the multimodal study of comedic timing. *arXiv preprint arXiv:2603.21803*, 2026.

A Author contributions

- Conceptualization of narrative categories: DB, KC, AC, JH, RK, MM, AP, INS, YS
- Benchmark question annotation: JH, RK, MM, AP, INS, YS (major); DB (minor)
- Benchmark answer annotation (evaluation): DB, KC, AC, JH, RK, MM, AP, RS, INS, YS
- *Variety* box office annotation: MM, DB
- Data curation (*Variety*, *Catalog of Copyright Entries*, Internet Archive): DB
- Statistical analysis: DB, KC
- Computational methodology: KC, DB
- Validation: DB, KC
- Writing: DB, AC, KC, RS

B Box office extraction accuracy

Our procedure for generating ranked lists of movies per year by their total box office numbers involves using a multimodal LLM to extract individual $\langle \text{city, theater, movie, gross} \rangle$ tuples from pages of *Variety* magazine, and map movie titles to IMDB identifiers (as described in the main body above). We evaluate the accuracy of this process by comparing the extractions derived from several models with those manually created by visually inspecting 12 issues of *Variety*.

In order to most closely measure the performance on our final goal (identifying the top n movies by box office per year), we apply the same process of mapping titles to IMDB identifiers for both the gold and extracted sets. If a movie does not exist in that mapping, we identify it by its lowercased title. We sum all gross values for the same movie across all tuples to yield a $\langle \text{movie, \$} \rangle$ set and generate a ranking over movies by their total gross. We do so over the gold annotations and the predicted values to generate two ranked lists.

We measure accuracy with three metrics:

1. For all movies that exist in the gold *and* predicted sets, we measure the Spearman rank correlation coefficient ρ between the gross dollar amounts in the two lists. This captures the degree to which the predicted ranks correspond with the true ranks, when we have a gross dollar amount for all movies.
2. To capture the degree to which the predicted ranks invent movies that do not exist, or fail to extract ones that do, we report the “Movie F1” score, where precision is defined as the fraction of predicted movies that exist in gold annotations and recall as the fraction of movies in the gold annotations that exist in the predicted set. Low precision would mean that models are hallucinating movies that do not exist; low recall would mean that models are failing to identify the ones that do.
3. While the two metrics above evaluate our final goal (assessing the ranking of movie titles by reported gross), we also evaluate the accuracy of exact tuples (i.e., the degree to which a tuple gets the movie title, theater, city and gross exactly correct).

Table 3 reports those metrics over several commercial model variants; each number represents the average of that metric over all *Variety* issues (so that each issue carries equal weight in assessment).

C Sample registrations

Figure 3 illustrates a copyright registration notice for the movie *The Bells of St. Mary’s* take from the *Catalog of Copyright Entries, Cumulative Series 1940–1949*; figure 4 shows the copyright registration renewal for that same film exactly 28 years later. In this case, the renewal references the registration date (6Dec45) and number (L81) of the original.

We use Gemini Pro 2.5 to extract the movie title, registration identifier and copyright date from each entry from the “Registrations” section of the CCE—e.g., $\langle \text{Bells of St. Mary’s, LP81, 6Dec45} \rangle$ —and movie title, original registration number, original copyright date, and renewal number and renewal

Table 3: Box office extraction accuracy, with 95% bootstrap confidence intervals.

Model	ρ	Movie F1	Tuple F1
Gemini 3 Pro	0.966 [0.953-0.977]	0.921 [0.909-0.934]	0.825 [0.800-0.851]
- (high \rightarrow low thinking level)	0.934 [0.912-0.950]	0.858 [0.802-0.896]	0.711 [0.636-0.770]
- (ultrahigh \rightarrow high res)	0.916 [0.887-0.938]	0.847 [0.800-0.886]	0.609 [0.533-0.680]
- (pro \rightarrow flash)	0.930 [0.903-0.951]	0.858 [0.826-0.886]	0.675 [0.633-0.728]
GPT 5.4 (original res)	0.916 [0.893-0.937]	0.809 [0.774-0.845]	0.684 [0.645-0.725]
- (mini)	0.605 [0.493-0.699]	0.388 [0.329-0.451]	0.083 [0.048-0.128]

THE BELLS OF ST. MARY'S, Rainbow Productions, Inc., c1945. 126 min., sd.
 Credits: Direction and story, Leo McCarey; screenplay, Dudley Nichols; music score, Robert Emmett Dolan; editor, Harry Marker.
 © Rainbow Productions, Inc.; 6Dec45; LP81.

Figure 3: Registration for *Bells of St. Mary's* recorded in the *Catalog of Copyright Entries, Cumulative Series 1940–1949*

BELLS OF ST. MARY'S, a photoplay in 14 reels by Rainbow Productions.
 © 6Dec45; L81. National Telefilm Associates, Inc. (PWH); 3Jan73; R542408.

Figure 4: Registration renewal for *Bells of St. Mary's* recorded in the *Catalog of Copyright Entries, Third Series, Volume 27, Parts 12–13, Number 1: Motion Pictures (January–June 1973)*

date from the “Renewals” section—e.g., (Bells of St. Mary’s, LP81, 6Dec45, R542408, 3Jan73) for volumes from 1957–1978 (corresponding to an earliest original registration date of 1929). We further extracted all motion picture renewals from the Copyright Public Records System to capture renewals made after 1978.

D Memorization

As noted in the main text, to identify questions that are highly memorized by models (answerable without access to the movie), we pass all questions through three frontier LLMs—Gemini Pro 3.1, Claude Opus 4.7 and GPT 5.5—with the following prompt: “Based on your knowledge of the movie {MOVIE} ({YEAR}), answer the following question.” All models exhibit similar rates of memorization (Gemini 39.4%, Opus 40.7%, GPT 38.8%). To mitigate this effect, we subselect questions from the benchmark pool so that the average performance across all models when prompted with the movie title and date alone is approximately 25% (reflecting a true random guess). The selection process ranks all questions in ascending order by the total number of the three models that correctly answer it from metadata alone, and adds questions sequentially to the benchmark until an average accuracy of 25% is reached. This yields a total of 628 benchmark questions (discarding 151 from the original pool). As Table 4 illustrates, the exclusion rate varies widely by category. For convenience, we let \mathcal{D} denote the full pool of 779 candidate questions, and \mathcal{B} the post-filter benchmark of 628 questions used in the experiments described §5.

Table 4: Exclusion rate by category.

Category	Exclusion rate	\mathcal{B} count	\mathcal{D} count
Representation	0.389	22	36
Symbolism	0.241	22	29
Temporality	0.221	53	68
Plot	0.209	155	196
Object Identification	0.200	92	115
Character	0.167	135	162
Perspective	0.155	49	58
Setting	0.130	100	115
Total		628	779

E Frame budget sweep

For the six end-to-end backbones run at multiple frame budgets (GPT-5-mini, Qwen3-VL-8B, and GLM-4.1V-9B-Thinking on the vision–language side; HourLLaVA, VAMBA-Qwen2-VL-7B, and VideoChat-Flash on the long-video side), average accuracy is reported in Table 5 (appendix E) for $N \in \{64, 128, 256\}$. The largest within-backbone gain over the full sweep is GPT-5-mini’s +3.6 pp from $N=64$ to $N=256$; the next largest is Qwen3-VL-8B’s +4.5 pp from $N=64$ to $N=128$. Within-backbone CIs overlap heavily across the three budgets, and three of the six backbones (Qwen3-VL-8B, GLM-4.1V-9B-Thinking, VideoChat-Flash, VAMBA-Qwen2-VL-7B in the non-monotone direction) do not improve monotonically with N . The frame-budget effect is an order of magnitude smaller than the audio-access gap shown in §5.2, and we therefore use a single canonical budget per sub-paradigm in Table 1 ($N=64$ for long-video models, $N=256$ for vision–language models).

Table 5: Frame-budget sweep $N \in \{64, 128, 256\}$ for end-to-end vision-language and long-video backbones run at multiple input budgets.

Backbone	$N=64$	$N=128$	$N=256$
<i>End-to-end vision-language models</i>			
GPT-5-mini	30.6 [27.2–33.9]	32.8 [29.3–36.3]	34.2 [30.7–37.7]
Qwen3-VL-8B	25.3 [22.0–28.5]	29.8 [26.4–33.1]	28.8 [25.5–32.2]
GLM-4.1V-9B-Thinking	25.8 [22.6–29.0]	25.6 [22.5–28.8]	24.2 [21.2–27.4]
<i>End-to-end long-video models</i>			
HourLLaVA	23.6 [20.4–26.8]	24.7 [21.5–27.9]	26.3 [23.1–29.5]
VAMBA-Qwen2-VL-7B	23.7 [20.5–26.9]	25.2 [22.0–28.5]	23.2 [20.1–26.4]
VideoChat-Flash	27.4 [24.0–30.7]	27.2 [23.9–30.6]	27.1 [23.7–30.4]

F Movies in Benchmark

Title	Year	IMDB Genres	Director	Production Company
McLintock!	1963	Comedy, Western	Andrew V. McLaglen	Batjac Productions
Beneath the 12-Mile Reef	1953	Adventure, Drama, Romance	Robert D. Webb	20th Century Fox
Cyrano de Bergerac	1951	Adventure, Drama, Romance	Michael Gordon	Stanley Kramer Productions
Go for Broke!	1951	Drama, History, War	Robert Pirosch	Loew’s
Royal Wedding	1951	Comedy, Musical, Romance	Stanley Donen	Loew’s

Continued on next page

(continued from previous page)

Title	Year	IMDB Genres	Director	Production Company
Three Guys Named Mike	1951	Comedy, Romance	Charles Walters	Metro-Goldwyn-Mayer (MGM)
The Inspector General	1949	Comedy, Musical, Romance	Henry Koster	Warner Bros.
Tulsa	1949	Drama, Western	Stuart Heisler	Walter Wanger Productions
He Walked by Night	1948	Crime, Drama, Film-Noir	Alfred L. Werker	Bryan Foy Productions
My Favorite Brunette	1947	Comedy, Crime, Mystery	Elliott Nugent	Hope Enterprises
The Perils of Pauline	1947	Drama, Romance	George Marshall	Paramount Pictures
Smash Up: The Story of a Woman	1947	Comedy, Crime, Drama	Stuart Heisler	Walter Wanger Productions
Till the Clouds Roll By	1947	Biography, Musical	Richard Whorf	Metro-Goldwyn-Mayer (MGM)
Angel on My Shoulder	1946	Adventure, Comedy, Fantasy	Archie Mayo	Charles R. Rogers Productions
The Strange Love of Martha Ivers	1946	Drama, Film-Noir, Romance	Lewis Milestone	Hal Wallis Productions
The Stranger	1946	Crime, Drama, Film-Noir	Orson Welles	International Pictures, The Haig Corporation
Blood on the Sun	1945	Drama, Romance, Thriller	Frank Lloyd	William Cagney Productions
Captain Kidd	1945	Adventure, Biography, Drama	Rowland V. Lee	Benedict Bogeaus Production
The Stork Club	1945	Comedy, Musical, Romance	Hal Walker	B.G. DeSylva Productions Inc.
Stage Door Canteen	1943	Comedy, Music, Romance	Frank Borzage	Sol Lesser Productions
Rudyard Kipling's Jungle Book	1942	Action, Adventure, Family	Zoltan Korda	Alexander Korda Films
Billy the Kid in Santa Fe	1941	Drama, Western	Sam Newfield	Sigmund Neufeld Productions
Meet John Doe	1941	Comedy, Drama, Romance	Frank Capra	Frank Capra Productions
Second Chorus	1941	Comedy, Musical, Romance	H.C. Potter	Boris Morros Productions
His Girl Friday	1940	Comedy, Drama, Romance	Howard Hawks	Columbia Pictures
Gulliver's Travels	1939	Adventure, Animation, Comedy	Dave Fleischer	Fleischer Studios
The Little Princess	1939	Comedy, Drama, Family	Walter Lang	20th Century Fox
Love Affair	1939	Comedy, Drama, Romance	Leo McCarey	RKO Radio Pictures
Made for Each Other	1939	Comedy, Drama, Romance	John Cromwell	Selznick International Pictures
Nurse Edith Cavell	1939	Biography, Drama, War	Herbert Wilcox	Imperadio Pictures Ltd.
Letter of Introduction	1938	Comedy, Drama, Mystery	John M. Stahl	Universal Pictures
A Star Is Born	1937	Drama, Romance	William A. Wellman	Selznick International Pictures
Swing High, Swing Low	1937	Comedy, Drama, Musical	Mitchell Leisen	Paramount Pictures

Continued on next page

(continued from previous page)

Title	Year	IMDB Genres	Director	Production Company
Little Lord Fauntleroy	1936	Drama, Family	John Cromwell	Selznick International Pictures
Becky Sharp	1935	Drama, Romance, War	Rouben Mamoulian	Pioneer Pictures Corporation
Of Human Bondage	1934	Drama, Film-Noir, Romance	John Cromwell	RKO Radio Pictures
A Farewell to Arms	1932	Drama, Romance, War	Frank Borzage	Paramount Pictures
Bird of Paradise	1932	Adventure, Drama, Romance	King Vidor	RKO Radio Pictures
Rain	1932	Drama	Lewis Milestone	Feature Productions
The Front Page	1931	Comedy, Crime, Drama	Lewis Milestone	The Caddo Company
Parlor, Bedroom and Bath	1931	Comedy	Edward Sedgwick	Metro-Goldwyn-Mayer (MGM)
Street Scene	1931	Drama, Romance	King Vidor	The Samuel Goldwyn Company, Feature Productions
All Quiet on the Western Front	1930	Drama, War	Lewis Milestone	Universal Pictures
Animal Crackers	1930	Comedy, Family, Musical	Victor Heerman	Paramount Pictures
Anybody's Woman	1930	Drama, Romance	Dorothy Arzner	Paramount Pictures
The Big House	1930	Crime, Drama, Thriller	George W. Hill	Metro-Goldwyn-Mayer (MGM), Cosmopolitan Productions
The Big Trail	1930	Adventure, Drama, Romance	Raoul Walsh	Fox Film Corporation
Check and Double Check	1930	Comedy	Melville W. Brown	RKO Radio Pictures
The Divorcee	1930	Drama, Romance	Robert Z. Leonard	Metro-Goldwyn-Mayer (MGM)
Feet First	1930	Adventure, Comedy, Family	Clyde Bruckman	The Harold Lloyd Corporation
Min and Bill	1930	Comedy, Drama	George W. Hill	Metro-Goldwyn-Mayer (MGM)
Reaching for the Moon	1930	Comedy, Romance	Edmund Goulding	Feature Productions
Song o' My Heart	1930	Drama, Music, Romance	Frank Borzage	Fox Film Corporation
Bulldog Drummond	1929	Crime, Drama, Mystery	F. Richard Jones	The Samuel Goldwyn Company
The Canary Murder Case	1929	Crime, Drama, Mystery	Malcolm St. Clair	Paramount Pictures
Coquette	1929	Drama, Romance	Sam Taylor	Pickford Corporation
Happy Days	1929	Comedy, Musical, Romance	Benjamin Stoloff	Fox Film Corporation
Sally	1929	Musical	John Francis Dillon	First National Pictures
The Trespasser	1929	Drama, Romance	Edmund Goulding	Gloria Productions
Weary River	1929	Drama, Romance	Frank Lloyd	First National Pictures
The Wild Party	1929	Comedy, Drama, Romance	Dorothy Arzner	Paramount Pictures