WHAT, WHEN, AND HOW: AUDIO DESCRIPTION AS CONSTRAINED GLOBAL OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio Description (AD) makes movies accessible to blind and visually impaired audiences by narrating visual information in gaps between dialogue. Existing automatic AD systems largely treat generation as a local video-to-text problem, assuming that the content to describe and its temporal location are already provided. Realistic AD instead requires coupled decisions about *what* visual information is narratively important, *when* it can be spoken without interfering with dialogue, and *how* it should be formulated to fit within the available time. We formalize AD generation as a constrained optimization problem over these three decisions. Our hybrid system uses large language models to propose and ground visual elements, estimate their salience to the narrative, and generate compressed realizations. A mixed-integer linear program then jointly selects and schedules descriptions across a scene subject to temporal constraints. When evaluated on REFRAMED, a benchmark for realistic AD of movies, our approach makes better decisions than prompted LLMs about what to describe and when to describe it, establishing a new SOTA on narrative QA and temporally grounded metrics. Ablations show that explicit temporal constraints drive gains in placement, while salience estimation controls how much narratively useful content is retained. Improvements are concentrated on temporal and narrative measures rather than n-gram overlap, although a significant gap to professional describers remains.

1 INTRODUCTION

Audio Description (AD) is a verbal narration of the key visual content of a movie, delivered in the gaps between dialogue so that blind and visually impaired audiences can follow the story. The central challenge of AD generation is deciding what visual information is most relevant, and what information can be left out (Vercauteren, 2016). A movie records countless details its director never deliberately chose (the colour of a passing car, the weather, what an extra is wearing), which is in contrast to a verbal narrative which contains only the details its narrator selected and asserted (Chatman, 1980). Describing a movie therefore means recovering the small number of visual details that the director intended an audience to read as story, and omitting the overwhelming majority that merely happen to be in frame. Vercauteren (2007) characterizes the describer's craft as a set of interacting decisions about the *what*, *when*, and *how* of narration: which visual elements carry the narrative, at what moment they can be spoken, and in what words.

The need to automate this work is now acute. In the US, Title II of the Americans with Disabilities Act will require public entities to provide AD for pre-recorded video, and the UK's Media Act sets a streaming quota of 10% by 2030, a volume that fully human description, which is slow and costly, cannot meet alone. Computational work has so far addressed largely only the last of the three decisions, reducing the task to a form of video captioning: the temporal placement of each description is taken from the reference AD, a clip is cut around it, and a model is trained or prompted to caption that clip, optionally conditioned on character information, screenplay context, or professional guidelines (Han et al., 2023a;b; Xie et al., 2024; Park et al., 2025; Li et al., 2025). This is a reasonable proxy for *how* a description should be phrased, and multimodal LLMs are now strong at describing an input video (Alayrac et al., 2022; Liu et al., 2023; Garg et al., 2024; Chai et al., 2025; Qwen Team, 2026b). However, it bypasses the problem of *what* to describe and *when* to describe it, by providing these as input to the model.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

[2s] *Lisbeth is in an upmarket tailors.* [3s] *The tailor shows her a leather motorcycle jacket he's made from a photo of Mikael wearing an identical garment.*

– [10s] **[SALANDER] It's nice.** – [12s] **[TAILOR] Your father?**
– [13s] **[SALANDER] A friend.** – [14s] **[TAILOR] Must be a very good friend.**

[15s] *She looks surprised by the tailor's comment.*

[18s] *Back at her apartment, she writes a card which has a picture of a horse on it* [22s] *and put it in an envelope.* [32s] *She addresses the envelope to 'M'.*

[35s] *Night.* [35s] *Lisbeth rides her motorcycle down a cobbled street with snow piled at the curbside.* [45s] *She parks on a corner outside Mikael's apartment* [48s] *and removes her crash helmet.* [50s] *She takes a package from the back of the bike,* [52s] *but stops dead when she sees Mikael leave his apartment with Erika.*

– [58s] **[MIKAEL] We're late.** – [60s] **[ERIKA] Fashionably late.**

[60s] *She watches numbly as the couple walk away with their arms around each other.* [78s] *Lisbeth looks down as Mikael and Erika get into a waiting taxi.* [87s] *Lisbeth turns away and flings the package which has the card attached to it into a nearby dumpster.* [93s] *She puts her crash helmet back on,* [95s] *gets on her bike* [96s] *and starts it up.* [97s] *A solitary figure, she rides off into the night.*

Figure 1: Excerpt of Audio Description (*italics*) and dialogue (**bold**) from *The Girl with the Dragon Tattoo* (2011); timestamps mark the start of narration.

Figure 1 illustrates what is lost when AD is reduced to captioning a pre-selected video clip. Consider *what* the describer selects from the closing scenes of *The Girl with the Dragon Tattoo* (see video here): a leather jacket made for Mikael, a card addressed to "M", and, eighty seconds later, that card attached to a package thrown into a dumpster. These details are selected because together they carry the narrative: Lisbeth has prepared a gift for Mikael, then discards it after seeing him leave with Erika, while countless co-occurring visual details are omitted. Consider *when* the descriptions are delivered: they occupy gaps in the dialogue and remain close to the visuals they describe, but need not be synchronous with them. The jacket, for example, is described before the exchange that refers to it, allowing the subsequent dialogue to be understood in context. Consider *how* the content is formulated: descriptions are concise to fit the available gaps, while still conveying narrative meaning, from the laconic "Night" to the more evocative "A solitary figure, she rides off into the night."

These decisions cannot be made independently, because they compete for narration time. Dense dialogue leaves little space, forcing a choice between describing fewer elements, describing them more tersely, or displacing a description away from the moment it refers to, by up to ten seconds (Sterner et al., 2026a). A natural solution is to learn a model that emits descriptions and timestamps jointly, but Sterner et al. (2026a) show that LLMs asked to do so produce descriptions of the wrong length for the gap they are given, overlapping with dialogue, or far from the visuals they describe. These are hard numerical constraints and LLM decoding provides no natural mechanism for enforcing them.

We therefore make the control explicit: neural models describe what happens on screen and when, and interpret it, while decisions that couple events under hard constraints are left to a solver. Our approach proceeds in three stages. First, a multimodal LLM describes the scene, and the description is segmented into events, each grounded to the span of video in which it occurs. Second, each event is scored for *narrative salience*: how much a viewer's understanding of the story would suffer were it omitted. Each event's description is also compressed to several shorter lengths, so that the same content is available both as a full sentence and as a shorter variant. Third, a solver decides across the whole scene at once which events to describe, which verbalization of each to use, and when to deliver it, maximizing total salience subject to hard constraints: narration must fall within a dialogue gap, stay close to the event it describes, and not overlap other narration. Because selection is discrete while delivery time is continuous, this is a mixed-integer linear program (MILP), for which mature and highly efficient solvers exist.

We evaluate on the REFRAMED benchmark (Sterner et al., 2026a), whose challenge set provides dual human-authored AD references and complementary measures of content, temporal alignment, and narrative comprehension. Across Qwen- and Gemini-based pipelines, constrained optimization yields its largest gains on the temporal and QA-based measures. The MILP also consistently outperforms an LLM scheduler given the same candidate events, occurrence spans, salience scores, narration durations, and dialogue gaps, demonstrating the value of explicit optimization. Our strongest system achieves the best automatic QA-based performance, although a substantial gap to human describers remains. Our contributions[1] are as follows:

[1]We will release all code necessary to reproduce our experimental results.

- A formalization of realistic AD generation as a constrained optimization problem that jointly decides *what* to describe, *when* to describe it, and *how* to formulate it, bringing together decisions that prior computational approaches optimize only partially or separately.
- A hybrid LLM-optimization system in which LLMs describe, ground, score, and compress events, and a mixed-integer linear program selects what to narrate, in which form, and when.
- Results on REFRAMED showing that constrained optimization outperforms an LLM making the same decisions from the same inputs, and conveys more of the story than prompted LLMs and captioning systems that are told when to describe.

2 Related Work

Most work on AD generation assumes that the temporal placement of each description is given by the reference AD. This reduces the task to video captioning: a model describes visually salient content of a pre-specified clip. Most systems fine-tune multimodal models for this task, augmenting inputs with signals such as broader visual context or character information (Han et al., 2023a;b; 2024; Lin et al., 2024; Deganutti et al., 2025; Wang et al., 2025; Ye et al., 2025), while others prompt LLMs zero-shot (Chu et al., 2024; Zhang et al., 2024; Xie et al., 2024). Despite differences in architecture and conditioning information, these approaches assume away the decision of *when* to describe and, by fixing the clip from the reference AD, strongly constrain *what* should be described.

A smaller body of work relaxes this assumption. Pavel et al. (2020) use dynamic programming to place human-scripted descriptions within audio gaps, shortening the text or lengthening the source audio where needed. Wang et al. (2021) predict insertion times from audiovisual inconsistency and select a description from a candidate set at each predicted time. Gupta et al. (2026) jointly predict insertion times and grounded visual content, then generate a description for those visuals, while Khandelwal et al. (2025) optimize content for predetermined temporal slots. These approaches address different subsets of the *what*, *when*, and *how* decisions, but none optimizes all three jointly across a scene: Pavel et al. (2020) assume human-written content; Wang et al. (2021) and Gupta et al. (2026) generate descriptions independently at predicted insertion points rather than allocating a shared narration budget; and Khandelwal et al. (2025) fix temporal placement.

Sterner et al. (2026a) formulate realistic AD generation as jointly deciding *what* to describe and *when*, and introduce REFRAMED, with dual human-authored references and metrics for this setting. REFRAMED provides the task and evaluation framework, but leaves open how these coupled decisions should be made under the hard temporal constraints imposed by the soundtrack. We extend this formulation by making *how*, the choice of wording and hence narration duration, an explicit decision, and optimize content selection, formulation, and placement jointly across the scene. The *what*/*when*/*how* decomposition itself comes from work on AD in translation studies (Vercauteren, 2007), where these decisions are guided by what the audience needs in order to follow the story (Vercauteren, 2016). These accounts are descriptive, however, and do not specify a computational mechanism for coordinating the three decisions.

Our formulation is closely related to global optimization approaches in text summarization, where salient content is selected under a limited linguistic budget. Early work casts summarization as global inference over relevance, redundancy, and length (McDonald, 2007), while integer-linear programming makes content selection explicit under a length constraint (Gillick et al., 2008; Gillick & Favre, 2009). Particularly relevant to our setting, such formulations can also choose among alternative compressed realizations of the same content (Madnani et al., 2007; Clarke & Lapata, 2008; Martins & Smith, 2009; Berg-Kirkpatrick et al., 2011), or jointly select and rewrite content under a shared budget (Woodsend & Lapata, 2010; 2012). We adopt the same separation between models that score candidate content and a solver that chooses among candidates. AD, however, replaces a single length budget with irregular temporal windows fixed by the soundtrack, while also requiring the optimizer to decide where within those windows selected content should be delivered.

3 Formalization

Building on the decisions illustrated in Figure 1, this section formalizes AD generation as decisions and constraints over *what* to say, *when* to say it, and *how* to say it. We first define at what level

decisions are made and introduce terminology. We then introduce decision variables for selecting, placing, and compressing descriptions. Finally, we state AD generation as an optimization problem.

Units and terminology. AD may describe events (something dynamic, such as Lisbeth parking), processes (something ongoing, such as Lisbeth riding her motorcycle), or states (something static, such as it being night-time). Following Bach (1986), we will refer to all of these as *eventualities*. Our formulation is as follows.

- Each scene depicts a number of eventualities $e_i \in \mathcal{E}$, obtained by segmenting a prose scene description and indexed in textual order.
- Each eventuality e_i has a candidate description element w_i.
- The eventuality is depicted in the video during an occurrence span $[\tau_i, \tau_i + \gamma_i]$.
- The available narration time may require a shorter formulation, so each element has $K + 1$ variants $\mathcal{C}_i = \{c_{i0}, \ldots, c_{iK}\}$, where $c_{i0} = w_i$ and c_{ik} is its k-th compression.
- If described, the eventuality is assigned a delivery span $[d_i, d_i + L_i]$, where L_i is the time required to narrate the variant used.

Occurrence and delivery spans are related but frequently non-equal: dialogue often prevents description at the moment an eventuality is on screen, in which case its description is delivered shortly before or after it.

Decision variables and further notation. There are three decision variables, reflecting the *what*, *when*, and *how* of AD.

- *What* — $x_i \in \{0, 1\}$ represents whether eventuality e_i is described or not.
- *When* — $d_i \in \mathbb{R}_{\geq 0}$ represents the delivery start time of the description of e_i.
- *How* — $y_{ik} \in \{0, 1\}$ represents whether variant c_{ik} is used.

Each eventuality is associated with a salience score $\sigma_i \in \mathbb{R}$, which represents its importance to the story being told. We define $l_{ik} = \delta(c_{ik})$, where δ maps text to narration duration, so that the narration time of the variant used is $L_i = \sum_{k=0}^{K} l_{ik} y_{ik}$. The mid-point of the occurrence span of e_i is $m_i = \tau_i + \frac{\gamma_i}{2}$. Finally, $\mathcal{G} = \bigcup_{j=1}^{J} [a_j, b_j]$ is the set of permissible narration times in a scene, formed by the union of J temporally disjoint closed intervals $[a_j, b_j]$.

AD as an optimization problem. Given candidate description elements and their variants, salience scores, occurrence spans, and permissible gaps, AD generation requires deciding which eventualities to describe, which variant to use for each, and when to deliver it. We state AD generation as the following optimization problem:

$$\underset{\mathbf{x}, \mathbf{y}, \mathbf{d}}{\arg\max} \quad \sum_i \sum_{k=0}^{K} \sigma_i \, l_{ik} \, y_{ik} \tag{1a}$$

$$\text{s.t.} \qquad \sum_{k=0}^{K} y_{ik} = x_i \qquad \forall i, \tag{1b}$$

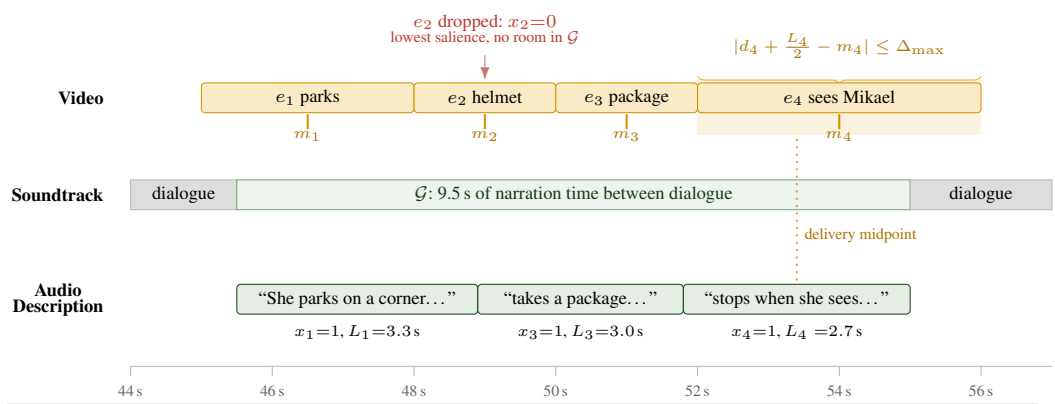
$$[d_i, d_i + L_i] \subseteq \mathcal{G} \qquad \forall i : x_i = 1, \tag{1c}$$

$$\left| d_i + \tfrac{L_i}{2} - m_i \right| \leq \Delta_{\text{max}} \qquad \forall i : x_i = 1, \tag{1d}$$

$$d_i + L_i \leq d_u \qquad \forall\, i < u : x_i = x_u = 1 \tag{1e}$$

The objective (1a) maximizes salience weighted by the narration time, so each second of a gap accrues the salience of the eventuality it describes. Within an eventuality this favours the longest variant that fits; without the weighting the solver would be indifferent between a full description and its shortest compression; across eventualities, a dense short description displaces a longer, less salient one whenever the time it frees can be filled at least as densely. The assumption that such content is available is met for AD: movies depict far more eventualities than can be narrated.

The objective is subject to four constraints. The first equation (1b) concerns compression: a described eventuality uses exactly one variant, and an omitted one uses none. The remaining three



The candidate wordings for e_4 are $\mathcal{C}_4 = \{c_{4,0}, \ldots, c_{4,3}\}$: the full description element $w_4 = c_{4,0}$ extracted from a scene description and three progressively shorter compressions of it. Exactly one is selected, by y_{4k}.

$c_{4,0}$	"but stops dead when she sees Mikael leave his apartment with Erika"	$l_{4,0} = 3.6\,\text{s}$	too long for $\mathcal{G}$
$c_{4,1}$	"but stops when she sees Mikael leave with Erika"	$l_{4,1} = 2.7\,\text{s}$	selected: $y_{4,1}{=}1$, so $L_4 = l_{4,1}$
$c_{4,2}$	"but stops when she sees Mikael and Erika"	$l_{4,2} = 2.4\,\text{s}$	fits, but loses *leaving*
$c_{4,3}$	"she stops dead"	$l_{4,3} = 0.9\,\text{s}$	fits, but loses who she sees

Figure 2: Example of the optimization, on the final scene of Figure 1. Four eventualities occur on screen (top), each with an occurrence midpoint m_i, while the soundtrack leaves one gap $\mathcal{G}$ between lines of dialogue (middle). The solution (bottom) makes all three decisions jointly: it sets $x_2{=}0$, dropping the least salient eventuality because the others cannot otherwise fit; it chooses delivery times d_i that tile $\mathcal{G}$ in order and without overlap; and it selects the compression $c_{4,1}$ for e_4: the full description $c_{4,0}$ would exceed the time left, and because the objective weights salience by narration time, the longest variant that fits scores highest. The shaded band shows the $\Delta_{\max}$ window that keeps a description near the event it describes. Durations and gap boundaries are illustrative.

concern temporal placement. They require each description to fall within a permissible narration interval (1c), to remain temporally close to the eventuality it describes (1d), and not to overlap the next description, which also preserves the order of the scene description (1e). Constraints conditioned on $x_i = 1$ are indicator constraints, and equation (1c) is a disjunction over intervals; Section 4 describes how both are linearized. The formulation is agnostic as to how its inputs are obtained; Section 4 describes our choices for scene description, occurrence spans, compression, and salience scoring.

AD Generation Example. Figure 2 illustrates the optimization on the final scene of Figure 1. Four eventualities occur on screen, while dialogue leaves a single gap of 9.5 seconds, which is too short to narrate all of them in full. The solver resolves this conflict through all three decisions at once. It drops the least salient eventuality, Lisbeth removing her helmet ($x_2 = 0$); it packs the remaining descriptions within the gap, in order and each within $\Delta_{\max}$ of its occurrence midpoint; and, since the full description of e_4 would overrun the gap, it selects a compression ($y_{4,1} = 1$), the longest variant that still fits. Appendix A walks through input preparation and the solution for a longer scene.

4 Experimental Setup

This section describes our data and evaluation protocol. We then describe how the MILP system is implemented, including its inputs, salience scoring alternatives, and the systems we compare against.

Data. We use REFRAMED (Sterner et al., 2026a), a dataset of 2,023 video excerpts (average length 144s) spanning 3,302 scenes from 206 movies. Each excerpt comes with two professional AD versions (American and British), professional dialogue subtitles, and, for 85 movies, screenplays aligned at the scene level. We report main results on the REFRAMED challenge set: ten full-length movies, manually annotated with scene boundaries and with two to three professionally transcribed AD versions each. The MILP is solved per scene; there are a total of 1,221 scenes, mean length is 57.4s. Ablations use a separate validation set of eight movies from the REFRAMED training and

validation splits, chosen as those whose screenplays best match the final post-production movie.[2] Screenplay scene descriptions are written to guide production, so they emphasize narrative intent rather than details fixed at the time of shooting, which makes it interesting to compare MILP inputs derived from screenplays against LLM-generated scene descriptions.

Evaluation. We follow the REFRAMED evaluation protocol, which compares generations against multiple references with three groups of metrics (details in Appendix B). *Dialogue-gap* metrics compute CIDEr and METEOR between generated and reference descriptions assigned to each gap in the dialogue subtitles. *Alignment* metrics use SODA (Fujita et al., 2020) to align generated and reference descriptions monotonically: SODA-M is the METEOR score of aligned pairs, and SODA-T the proportion of references aligned within $\tau = 10$ seconds. *QA-based* metrics measure narrative usefulness: QEval is the proportion of reference-based questions generated AD answers correctly; QEval-T requires the answer comes from a description within τ seconds of the correct time. Statistical significance is assessed with two-tailed Monte Carlo permutation tests ($R = 10{,}000$, $\alpha = 0.05$).

Prompted LLMs can be rewarded for descriptions that in practice cannot be narrated. We therefore also report **realistic** variants (the indented rows in Table 1), which remove descriptions that would need to be narrated faster than 300 WPM,[3] fall outside a dialogue gap (with a one-second collar), or begin before the preceding description ends. The filter applies only to the prompted LLMs. Our MILP satisfies these conditions by construction: constraint equation 1c places every description inside a dialogue gap, constraint equation 1e prevents descriptions from overlapping, and narration durations are computed at 200 WPM, so no description is narrated faster than it can be spoken.

MILP system. We build two versions of the system, in which all MILP inputs are produced by either Qwen 3.5 27B (Qwen Team, 2026a) or Gemini 3.1 Flash-Lite (Gemini Team, 2026). Video is sampled at one frame per second, without audio. The Qwen-based pipeline runs locally with vLLM on two H200 GPUs; Gemini is accessed through the official paid API.

The LLM generates a prose scene description from the video (prompt in Appendix C.1). Because video input alone does not support correct character naming, we augment it with two sources of automatically computed character information: character names with face crops (our IMDb-portrait comparison method achieves $F_1 = 73.7$; Appendix C.2.3), and dialogue timestamps labeled with the speaking character (our speaker diarization method achieves $F_1 = 77.9$; Appendix C.2.1). The scene description is segmented into sentences (Frohmann et al., 2024) and then into description elements using REFRAMED's learned segmentation model (character-level $F_1 = 66.0$ vs. 38.4 for a comma-splitting baseline). A second prompt (Appendix D) takes the video and the full sequence of description elements, and returns for each element its occurrence span and five compressed variants, at 0.9, 0.8, 0.7, 0.6 and 0.5 times the source length. The same prompt also produces our default salience scores, assigning each element a score in $[0, 1]$; salience represents the extent to which the element provides information necessary for following the story, relative to the other elements of the scene. Each variant is assigned a narration duration at a fixed speaking rate of 200 words per minute, computed from the generated words. Permissible narration times $\mathcal{G}$ are the gaps of at least one second between REFRAMED dialogue subtitles.

Because salience drives content selection, we compare our LLM scores against four alternatives and two oracles, leaving the rest of the pipeline unchanged. **Random** draws each score uniformly from $[0, 1]$. The other three follow Sterner et al. (2026b), who define salience as the similarity between an element and a representation of the narrative as a whole. **Embedding video** and **Embedding text** take the cosine similarity between the element embedding and an embedding of the full video and of the full scene description, respectively, using Qwen 3 VL Embedding (8B) over the Qwen descriptions; Video embeddings are native, with the video-trained model encoding all frames jointly. **BM25** scores the lexical overlap between the element and the scene description. The two oracles indicate how much headroom better salience estimation can contribute, by scoring each element against the reference AD: **BM25 oracle** and **Embedding oracle** use BM25 and embedding similarity, respectively. Salience scores are not normalized; empirically, we find all predictions are non-negative.

[2] For our method for scoring the degree of match and an analysis of the correlation of this measure against a QA-based evaluation of the screenplay scene descriptions, see Appendix C.3.

[3] The LLM prompts request 180 WPM, the rate in American AD guidelines (Audio Description Coalition, 2009); over 99% of REFRAMED reference descriptions are delivered below 300 WPM.

The MILP is solved with Gurobi 13.0.3 (Gurobi Optimization, LLC, 2026) on a node with 24 CPUs. Gap containment equation 1c is linearized with binary gap-assignment variables, and the constraints conditioned on $x_i = 1$ are encoded as Gurobi indicator constraints. We set $\Delta_{\max} = 10$ seconds, following the analysis of manually timecoded AD in Sterner et al. (2026a), which shows that professional describers narrate an element within ten seconds of its occurrence (in 99.4% of cases). Solving stops when Gurobi's default optimality criteria are met or after ten minutes, whichever comes first. Of the scenes solved optimally, mean and median solve times are 2.75s and 0.12s, respectively. The solving for 11 and 15 scenes was stopped due to the ten-minute timeout, for Qwen and Gemini inputs respectively; these have a mean optimality gap of 18.1% and 19.2%. The conditions that lead to timeout are a complex combination of the objective and constraints: for instance, one scene with Qwen inputs reaches the timeout despite only being 32 seconds long, a result of the 33 candidate eventualities needing to be scheduled into only two dialogue gaps of total length 14.8s.

Comparison systems. We compare against the published generations and results of Sterner et al. (2026a), reporting their best variant under the evaluation protocol: **Qwen 3.5** and **Gemini 3.1**, the same models used to produce the MILP inputs, here prompted with the video to generate timecoded descriptions directly. Their model is given the video together with the dialogue subtitles and the list of gaps of at least one second, and is not told which gaps should contain a description. The lower bound is a greedy **Random** baseline that fills each dialogue gap with randomly sampled training-set description elements until no further element fits. The upper bound is an **Expert** AD transcript, evaluated as a system output. **ShotbyShot** (Xie et al., 2025) and **DistinctAD** (Fang et al., 2025) are specialist AD systems that generate one description sentence from one clip and its context.

Finally, a controlled comparison replaces the solver with an LLM. The **Qwen scheduler** and **Gemini scheduler** receive the same information as the MILP (description elements, occurrence spans, salience scores, 200 WPM narration durations, and dialogue gaps), together with the text of the elements and their compressions, which the MILP does not use. They produce the same outputs as the MILP, a selected variant and a delivery start time for each element, relying on the LLM's parametric knowledge of AD rather than the symbolic constraints given to the MILP. We run Qwen 3.5 with reasoning enabled and Gemini 3.1 with `thinking=high`, as in all other LLM runs.

5 Results and Discussion

Does the MILP reach state-of-the-art automatic AD? Table 1 reports results on the REFRAMED challenge set. Among automatic systems, Gemini MILP performs best on both QA-based metrics (QEval=45.9 and QEval-T=25.5; $p < 0.01$ against every other automatic system). It also achieves the best SODA-T score except for ShotbyShot, with which the difference is not significant. Qwen MILP shows the same pattern: its largest gains over realistic Qwen are on SODA-T (53.6 vs. 31.2) and QEval-T (22.1 vs. 14.8; both $p < 0.01$). DistinctAD and ShotbyShot are advantaged on SODA-T (49.0 and 53.0) because they inherit the reference AD's placement and content selection, but this does not translate into comparable narrative performance (QEval-T=8.8 and 17.0). MILP systems do not improve CIDEr (e.g., Gemini MILP 13.6 vs. Gemini 3.1 19.0), which is not unexpected: CIDEr rewards n-gram overlap with the reference wording, which favours the fluent full sentences the prompted LLMs produce, whereas the MILP's compression and selection yield terser, sometimes fragmentary descriptions (see Figure 3) that convey the same content in different words. The gain on QA-based metrics alongside the CIDEr drop reflects this: the MILP trades surface overlap with a single reference for narrative usefulness. Under all

Table 1: System performance on REFRAMED challenge set with six evaluation metrics. Indented rows apply the realistic filter. (Dialogue) gap columns are CIDEr and METEOR; SODA columns are SODA-M and SODA-T; QEval columns are QEval (Acc) and QEval-T (T).

	Gaps		**SODA**		**QEval**	
	C	**M**	**M**	**T**	**Acc**	**T**
Expert	51.4	20.1	16.3	81.0	69.6	61.2
Random	1.3	4.4	4.0	27.1	34.1	2.3
DistinctAD	15.6	7.1	8.3	49.0	35.7	8.8
ShotbyShot	16.3	7.0	8.0	53.0	39.9	17.0
Qwen 3.5	13.2	8.1	7.4	38.4	43.9	17.3
realistic	11.3	6.5	7.1	31.2	42.6	14.8
Gemini 3.1	19.0	8.2	7.9	36.0	42.3	16.4
realistic	16.9	7.3	7.7	32.0	41.9	15.1
Qwen scheduler	14.0	7.1	7.2	47.6	43.3	21.8
Qwen MILP	11.4	7.9	7.7	53.6	43.6	22.1
Gemini scheduler	14.5	6.6	6.9	45.8	41.6	19.9
Gemini MILP	13.6	8.4	8.0	55.7	45.9	25.5

[HERMIONE] I won't be going alone, because, believe it or not, someone's asked me!

American AD: [+3s] *Hermione glares at Ron.* [+5s] *She hands her homework to Snape* [+6s] *and grabs her textbook from the table.*
British AD: [+4s] *She gets up,* [+5s] *hands her book to Snape,* [+6s] *and returns.*

Qwen 3.5: *[No AD]*
Qwen scheduler: [+6s] *She stands up.* [+9s] *Gives Snape papers.*
Qwen MILP: [+4s] *Snape walks down the aisle collecting papers.* [+6s] *Rises from seat.* [+7s] *Walks to teacher.* [+8s] *And hands him a stack of sheets.* [+10s] *Looks at note.*
Gemini 3.1: [+4s] *Ron and Hermione whisper during class.* [+7s] *Snape walks past them.*
Gemini scheduler: [+7s] *Snape stops at the table where Hermione sits.*
Gemini MILP: [+4s] *Snape stops at the table where Hermione sits.* [+6s] *She gives him a book.* [+8s] *Hermione focuses.* [+9s] *She then turns back to her own work.*

[+11s] **[HERMIONE] And I said yes!**

To whom does Hermione hand her homework? **A. Snape** B. George C. Harry D. Fred E. Ron

Figure 3: References and generated ADs for dialogue gap in *Harry Potter and the Goblet of Fire* (2005). Dialogue in **bold**, AD in *italics*. Multiple-choice question is from the QA-based evaluation.

evaluation metrics, the expert human upper bound significantly outperforms all other systems (all $p < 0.01$), and the random baseline is outperformed by all other systems (all $p < 0.01$).

Does explicit optimization outperform an LLM scheduler? Under the controlled comparison between the MILP and LLM schedulers, the MILP improves results for both Gemini and Qwen on all but the lexical metrics. The MILP is significantly better under every metric except CIDEr, on which it is indistinguishable from the Gemini scheduler and worse than the Qwen scheduler ($p < 0.01$); all other differences have $p < 0.05$. The gains are largest on temporal placement, and larger for Gemini than for Qwen: Gemini MILP improves on the Gemini scheduler under both temporal metrics (SODA-T 55.7 vs. 45.8; QEval-T 25.5 vs. 19.9), whereas Qwen differences are concentrated in SODA-T (53.6 vs. 47.6), with only numerically small gains on QEval-T (22.1 vs. 21.8). These results show that explicit optimization improves coordination of the *what*, *when*, and *how* decisions, while the scheduler's strong performance suggests that the staged representation is helpful.

On average the MILP selects 45.2% of candidate eventualities for Qwen and 41.2% for Gemini, and it selects fewer when narration time is scarcer. For Gemini, the share selected rises steadily with available narration time, from 21.0% when under 5s is free, through 32.7%, 48.7%, 67.2% and 80.6%, to 91.8% when more than 80s is available. Among the selected eventualities, 49.4% keep their full variant and the rest are compressed, with the compressions skewed toward lighter rates (0.9: 18.6%, 0.8: 9.7%, 0.7: 5.0%, 0.6: 6.2%, 0.5: 11.2%). This follows from the length-weighted objective, which favours the full variant when there is space but is forced toward heavier compression, and hence the tail at 0.5, when gaps are tight. Table 2 turns to how SODA-T and QEval vary with solve time, for both the MILP and the LLM scheduler. The MILP's advantage (Δ) grows as solve time increases, showing the solver's strength precisely when selection, placement, and compression must be tightly coordinated.

Table 2: Performance stratified by MILP solve time, an empirical proxy for *AD difficulty*. As scenes become harder to schedule (longer runtimes), scores drop, while the MILP's advantage (Δ) over the LLM scheduler widens.

	SODA-T			QEval		
Solve time	LLM	MILP	Δ	LLM	MILP	Δ
<0.1s	57.5	62.4	+4.9	48.3	51.4	+3.1
0.1–1s	46.9	57.0	+10.1	42.2	45.8	+3.6
1–10s	45.0	55.6	+10.6	43.7	48.3	+4.6
10–60s	37.8	50.1	+12.3	33.9	40.8	+6.9

Do LLMs game the AD metrics? The realistic variants (Table 1) are significantly worse than raw LLMs across all metrics (all $p < 0.01$). Dialogue overlaps are rare (1.6% and 1.2% of generations for Qwen 3.5 and Gemini 3.1) and no two descriptions overlap each other; the dominant failure is excessive speaking rate, with 28.9% and 16.7% of generations exceeding the 300 WPM ceiling, and a third and a tenth of those respectively exceeding even 500 WPM. These LLMs thus lack the temporal and linguistic control AD requires, and the realistic filter prevents them from exploiting the metrics with descriptions that could not be narrated. The MILP satisfies the filter by construction, so it never alters results; the LLM scheduler is likewise reported after filtering for fairness, with near-identical numbers (dialogue-gap scores are unchanged and all other differences are ≤ 0.4 points).

Consider Figure 3, which shows a seven second dialogue gap half way into a *Harry Potter* movie. Snape is walking the aisles of a classroom, as students work. Hermione reveals that someone has

asked her to the upcoming Christmas ball, then gets up, hands her homework to Snape, and returns to say that she has accepted. The prompted Qwen 3.5 does not generate anything for this dialogue gap, while Gemini describes the characters whispering and Snape walking past (which occurs much earlier). The Qwen scheduler is better, but using only six words is extremely terse, while the Gemini scheduler selects only a single 2 second description about Snape. Meanwhile, both MILPs describe Hermione giving Snape her homework, which is what the relevant evaluation QA pair tests. Note that as a result of the selection decisions in the MILP, the generations are not perfectly fluent, and the Qwen MILP uses an incomplete sentence (*Rises from seat*).

Which components contribute to performance? Table 3 ablates Qwen MILP on the validation set. Removing the grounding constraint (row 2) primarily hurts temporal performance (SODA-T 54.9 $\rightarrow$ 47.8; QEval-T 28.2 $\rightarrow$ 16.0), while removing compressed variants (row 3) reduces content-based metrics (CIDEr 12.9 $\rightarrow$ 9.2; QEval 48.4 $\rightarrow$ 43.4), because fewer descriptions can fit the available gaps. Removing both gives the worst overall performance (row 4). Thus, grounding determines where descriptions can be placed; compression increases how much useful content can be retained.

Table 3: Qwen performance on REFRAMED post-production screenplay set using **C**IDEr and **M**ETEOR, SODA-**M**, SODA-**T**, QEval (**Acc**), and QEval-**T**.

	Gaps		SODA		QEval	
	C	**M**	**M**	**T**	**Acc**	**T**
(1) Qwen MILP	12.9	7.7	7.3	54.9	48.4	28.2
MILP decisions and constraints						
(2) w/o grounding	11.0	7.0	7.4	47.8	44.7	16.0
(3) w/o compression	9.2	6.2	7.3	43.7	43.4	22.7
(4) + w/o both	6.6	6.0	7.6	41.0	39.7	12.9
Salience scoring						
(5) Random	12.1	7.2	7.1	53.1	43.6	20.0
(6) Embed. video	11.2	6.9	6.9	51.3	41.6	20.4
(7) Embed. text	12.3	7.3	7.3	50.7	43.8	22.5
(8) BM25	13.5	7.6	7.3	54.2	46.0	22.2
(9) BM25 oracle	16.3	8.6	8.0	51.4	46.6	23.8
(10) Embed. oracle	16.5	8.8	7.8	54.8	49.9	24.2
Scene descriptions						
(11) Screenplay	13.5	7.3	7.4	47.7	53.1	31.1

Rows (5)–(10) vary salience scoring, leaving the remaining pipeline fixed. Non-oracle methods differ little on lexical and temporal metrics, but separate more clearly on QA: LLM salience reaches QEval=48.4 and QEval-T=28.2, compared with 43.6 and 20.0 for random salience. Better salience therefore primarily improves *what* is selected rather than *when* it is delivered. The oracle scores support this: they improve CIDEr (16.3–16.5 vs. 12.9) and not SODA-T or QEval-T.

Finally, replacing the LLM scene description with screenplay descriptions (row 11) gives the best QA performance (QEval=53.1; QEval-T=31.1), suggesting screenplays better capture narrative intent. Temporal alignment falls (SODA-T 54.9 $\rightarrow$ 47.7), plausibly because pre-production screenplay events are harder to ground in the final video. Appendix F provides qualitative examples.

6 Conclusion

We formalized realistic AD generation as a constrained optimization problem over what to describe, when to deliver it, and how to formulate it. We instantiated this view in a hybrid system in which LLMs propose, ground, score, and compress visual events while a mixed-integer linear program makes joint decisions across a scene. On REFRAMED, the resulting system achieves the strongest QA-based performance among automatic systems and improves over an LLM scheduler given the same structured inputs. The gains are concentrated in temporal placement and narrative usefulness rather than n-gram overlap, reflecting the aspects of AD that the formulation explicitly controls.

Our analysis ablations clarify where these gains come from. Temporal grounding and compression are critical for placing descriptions within the limited narration time, while salience primarily determines which parts of the story are retained. Oracle salience scores reveal further headroom in content selection, although they do not consistently improve temporal placement or QA-based measures; screenplay-based scene descriptions, by contrast, yield the largest gains on QA-based measures, suggesting that representing narrative importance remains a major bottleneck. At the same time, a substantial gap to professional describers remains in both content selection and temporal realization. Closing it will require better models of narrative salience and visual grounding; importantly, these improvements can be incorporated as better inputs to the same global optimization framework.

AI USE STATEMENT

In this work, we used generative AI tools for assistance with coding implementation and for writing feedback. We have not used generative AI tools for any of the following: generating synthetic data (all reference AD used is human-authored), help with the conceptual development of our framework, formulation of mathematical claims, assistance in the writing of proofs, proposals of hypotheses, design of our research methodology or experiments, translation, data cleaning, or results interpretation. We have reviewed all AI-assisted work manually. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736. Curran Associates, Inc., 2022.

Audio Description Coalition. *Standards for Audio Description and Code of Professional Conduct for Describers based on the training and experience of audio describers and trainers from across the United States*, 3rd edition, 2009. URL https://audiodescriptionsolutions.com/wp-content/uploads/2020/04/adc_standards_090615.pdf.

Emmon Bach. The algebra of events. *Linguistics and Philosophy*, 9(1):5–16, 1986. ISSN 01650157, 15730549. URL http://www.jstor.org/stable/25001229.

Taylor Berg-Kirkpatrick, Dan Gillick, and Dan Klein. Jointly learning to extract and compress. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea (eds.), *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 481–490, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-1049/.

Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher D Manning. AuroraCap: Efficient, performant video detailed captioning and a new benchmark. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=tTDUrseRRU.

Seymour Chatman. What novels can do that films can't (and vice versa). *Critical Inquiry*, 7:121–140, 1980. URL https://doi.org/10.1086/448091.

Peng Chu, Jiang Wang, and Andre Abrantes. LLM-AD: Large language model based audio description system, 2024.

James Clarke and Mirella Lapata. Global inference for sentence compression: An integer linear programming approach. *Journal of Artificial Intelligence Research*, 31:399–429, 2008. doi: 10.1613/jair.2433.

Adrienne Deganutti, Simon Hadfield, and Andrew Gilbert. DANTE-AD: Dual-vision attention network for long-term audio description. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition - Workshop on AI for Content Creation (AICC'25)*, 2025.

Bo Fang, Wenhao Wu, Qiangqiang Wu, Yuxin Song, and Antoni B. Chan. DistinctAD: Distinctive audio description generation in contexts. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 13571–13581, June 2025.

Markus Frohmann, Igor Sterner, Ivan Vulić, Benjamin Minixhofer, and Markus Schedl. Segment Any Text: A universal approach for robust, efficient and adaptable sentence segmentation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11908–11941, Miami, Florida, USA,

November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.665. URL https://aclanthology.org/2024.emnlp-main.665/.

Soichiro Fujita, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. SODA: Story oriented dense video captioning evaluation framework. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 517–531, 2020. doi: 10.1007/978-3-030-58539-6_31.

Roopal Garg, Andrea Burns, Burcu Karagol Ayan, Yonatan Bitton, Ceslee Montgomery, Yasumasa Onoe, Andrew Bunner, Ranjay Krishna, Jason Michael Baldridge, and Radu Soricut. ImageInWords: Unlocking hyper-detailed image descriptions. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 93–127, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.6. URL https://aclanthology.org/2024.emnlp-main.6/.

Gemini Team. Gemini 3.1 Flash-Lite: Built for intelligence at scale, March 2026. URL https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-flash-lite/.

Dan Gillick and Benoit Favre. A scalable global model for summarization. In James Clarke and Sebastian Riedel (eds.), *Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*, pp. 10–18, Boulder, Colorado, June 2009. Association for Computational Linguistics. URL https://aclanthology.org/W09-1802/.

Dan Gillick, Benoit Favre, and Dilek Hakkani-Tür. The ICSI summarization system at TAC 2008. In *Proceedings of the Text Analysis Conference (TAC)*, 2008.

Akshita Gupta, Aditya Arora, Federico Tombari, Marcus Rohrbach, and Anna Rohrbach. From visual cues to spoken narration: Rethinking audio description. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2026. URL https://arxiv.org/abs/2609.01725.

Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2026. URL https://www.gurobi.com.

Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. AutoAD: Movie description in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18930–18940, June 2023a.

Tengda Han, Max Bain, Arsha Nagrani, Gul Varol, Weidi Xie, and Andrew Zisserman. AutoAD II: The sequel - who, when, and what in movie audio description. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13645–13655, October 2023b.

Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. AutoAD III: The prequel - back to the pixels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18164–18174, June 2024.

Eshika Khandelwal, Junyu Xie, Tengda Han, Max Bain, Arsha Nagrani, Andrew Zisserman, Gül Varol, and Makarand Tapaswi. More than a moment: Towards coherent sequences of audio descriptions, 2025. URL https://arxiv.org/abs/2510.25440.

Bruno Korbar, Jaesung Huh, and Andrew Zisserman. Look, listen and recognise: Character-aware audio-visual subtitling. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2975–2979. IEEE, 2024.

Anne Lambert, Marie Guégan, and Kai Zhou. Scene reordering in movie script alignment. In *2013 11th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pp. 213–218, 2013. doi: 10.1109/CBMI.2013.6576585.

Chaoyu Li, Sid Padmanabhuni, Maryam S Cheema, Hasti Seifi, and Pooyan Fazli. VideoA11y: Method and dataset for accessible video description. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–29, 2025.

Kevin Qinghong Lin, Pengchuan Zhang, Difei Gao, Xide Xia, Joya Chen, Ziteng Gao, Jinheng Xie, Xuhong Xiao, and Mike Zheng Shou. Learning video context as interleaved multimodal sequences. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 375–396, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023.

Nitin Madnani, David Zajic, Bonnie Dorr, Necip Fazil Ayan, and Jimmy Lin. Multiple alternative sentence compressions for automatic text summarization. In *Proceedings of the Document Understanding Conference at NLT/NAACL*, 2007.

André Martins and Noah A. Smith. Summarization with a joint model for sentence extraction and compression. In James Clarke and Sebastian Riedel (eds.), *Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*, pp. 1–9, Boulder, Colorado, June 2009. Association for Computational Linguistics. URL https://aclanthology.org/W09-1801/.

Ryan McDonald. A study of global inference algorithms in multi-document summarization. In Giambattista Amati, Claudio Carpineto, and Giovanni Romano (eds.), *Advances in Information Retrieval*, pp. 557–564, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-71496-5.

Andrew Cameron Morris, Viktoria Maier, and Phil D Green. From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition. In *Proceedings of Interspeech*, pp. 2765–2768, 2004.

Jaehyeong Park, Junchel Ye, Seungkook Lee, Hyun W. Ka, and Dongsu Han. NarrAD: Automatic generation of audio descriptions for movies with rich narrative context. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pp. 409–419, February 2025.

Amy Pavel, Gabriel Reyes, and Jeffrey P. Bigham. Rescribe: Authoring and automatically editing audio descriptions. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, UIST '20, pp. 747–759, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450375146. doi: 10.1145/3379337.3415864. URL https://doi.org/10.1145/3379337.3415864.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.5-Omni technical report, 2026b. URL https://arxiv.org/abs/2604.15804.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Rohit Saxena and Frank Keller. MovieSum: An abstractive summarization dataset for movie screenplays. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4043–4050, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.239. URL https://aclanthology.org/2024.findings-acl.239/.

Igor Sterner, Mirella Lapata, Alex Lascarides, and Frank Keller. REFRAMED: Towards realistic audio description generation for movies. In *Proceedings of the Conference on Language Modeling (COLM)*, October 2026a. URL https://doi.org/10.48550/arXiv.2608.09765.

Igor Sterner, Alex Lascarides, and Frank Keller. Contrastive learning with narrative twins for modeling story salience. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1528–1550, Rabat, Morocco, March 2026b. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.71. URL https://aclanthology.org/2026.eacl-long.71/.

Gert Vercauteren. Towards a european guideline for audio description. In Jorge Díaz Cintas, Pilar Orero, and Aline Remael (eds.), *Media for All: Subtitling for the Deaf, Audio Description, and Sign Language*, pp. 139–149. Brill, Leiden, The Netherlands, 2007. ISBN 9789401209564. doi: 10.1163/9789401209564_011. URL https://brill.com/view/book/9789401209564/B9789401209564-s011.xml.

Gert Vercauteren. *A narratological approach to content selection in audio description*. PhD thesis, Antwerp University, Belgium, 2016. URL https://repository.uantwerpen.be/docman/irua/4a8d3c/11347.pdf.

Hanlin Wang, Zhan Tong, Kecheng Zheng, Yujun Shen, and Limin Wang. Contextual AD narration with interleaved multimodal sequence. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 8372–8383, June 2025.

Yujia Wang, Wei Liang, Haikun Huang, Yongqi Zhang, Dingzeyu Li, and Lap-Fai Yu. Toward automatic audio description generation for accessible videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445347. URL https://doi.org/10.1145/3411764.3445347.

Kristian Woodsend and Mirella Lapata. Automatic generation of story highlights. In Jan Hajič, Sandra Carberry, Stephen Clark, and Joakim Nivre (eds.), *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 565–574, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL https://aclanthology.org/P10-1058/.

Kristian Woodsend and Mirella Lapata. Multiple aspect summarization using integer linear programming. In Jun'ichi Tsujii, James Henderson, and Marius Paşca (eds.), *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 233–243, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://aclanthology.org/D12-1022/.

Junyu Xie, Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. AutoAD-Zero: A training-free framework for zero-shot audio description. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pp. 2265–2281, December 2024.

Junyu Xie, Tengda Han, Max Bain, Arsha Nagrani, Eshika Khandelwal, Gül Varol, Weidi Xie, and Andrew Zisserman. Shot-by-Shot: Film-grammar-aware training-free audio description generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16503–16513, October 2025.

Xiaojun Ye, Chun Wang, Yiren Song, Sheng Zhou, Liangcheng Li, and Jiajun Bu. FocusedAD: Character-centric movie audio description, 2025. URL https://arxiv.org/abs/2504.12157.

Chaoyi Zhang, Kevin Lin, Zhengyuan Yang, Jianfeng Wang, Linjie Li, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. MM-Narrator: Narrating long-form videos with multimodal in-context learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13647–13657, June 2024.

A Detailed Worked Example

We use the tailor scene that opens Figure 1 to walk through the steps taken to prepare the inputs for and solve the AD optimization problem of Section 3. The input is the video.[4] The output is textual description elements, each with a start time for their delivery, as shown in Figure 1 (page 2). The starting point for solving the optimization problem is pre-determining the required inputs.

- **Permissible narration times.** The set of Permissible narration times are the gaps between character dialogue. In the example, character dialogue occurs in two temporal spans: in second 10–14, and in second 58–61. The Permissible narration times are the complement: second 0–10, 14–58, and 61–97 (the end of the video). Within the first scene, which ends at 18s, the permissible narration times are 0–10s and 14–18s.
- **Scene description.** The eventualities depicted in the first scene can be expressed as a prose scene description, see the descriptions in Table 4.
- **Description elements.** The scene description can be split up into the individual eventualities that are depicted. Using the scene description as input, this process can be seen as a form of text segmentation. The table shows the 15 description elements.
- **Occurrence spans.** Each eventuality is depicted at a time that establishes it for a sighted viewer. Temporal grounding provides this temporal information. This interval also permits computation of the midpoint of each occurrence span (the average of the printed numbers).
- **Description compression.** Each description is reformulated into $K = 5$ compressed variants at 0.9, 0.8, 0.7, 0.6 and 0.5 of its word count. For example, the first element *Lisbeth stands with her hands on a table in an upmarket Stockholm tailor shop.*, 14 words, could have variants *Lisbeth stands with her hands on a table in an upmarket tailor shop.*, 13 words; *Lisbeth leans on a table in an upmarket Stockholm tailor shop.*, 11; *Lisbeth leans on a table in an upmarket tailor shop.*, 10; *Lisbeth stands in an upmarket Stockholm tailor shop.*, 8; and *Lisbeth stands in an upmarket tailor shop.*, 7. Each description element is accordingly transformed into K compressions, resulting in $K + 1$ variants.
- **Compressed durations.** Each compressed variant is associated with a scalar that represents the duration required for narrating it.
- **Salience.** Each eventuality is also associated with a salience score, representing its relevance to the story being told. Illustrative salience scores are provided in the table.

With these pre-determined inputs, the optimizer operates over three decision variables. The final value for these decision variables also serves as the output of the optimization.

- $\mathbf{x} \in \{0, 1\}^{15}$ is a binary vector, with each element representing the selection (or rejection) of each of the 15 eventualities.
- $\mathbf{d} \in \mathbb{R}^{15}$ is a continuous vector, with each element representing the starting delivery time for its corresponding selected eventuality.
- $\mathbf{y} \in \{0, 1\}^{15\times(K+1)}$ is a binary matrix, with each element representing the selection of each compressed variant.

The total duration required to narrate the full scene description exceeds the time available between dialogue. The optimization resolves this conflict by selecting the subset of the inputs that maximizes the salience-based objective. Narrating the full scene description takes 36s at 200 WPM, and the scene leaves 14s between dialogue: 0–10s and 14–18s. With $\Delta_{\max} = 10$s, every eventuality can be placed in either gap, so the conflict must be resolved by selection and compression rather than by eligibility. The optimal solution selects five of the 15 eventualities. The first gap places e_1 at $d_1 = 0$s, e_{10} at $d_{10} = 4.0$s and e_{11} at $d_{11} = 6.1$s, ending at 10.0s; the second holds e_{14} at 14.1s and e_{15} at 16.2s, ending at 18.0s. In full, e_1, e_{10} and e_{11} would take 10.2s and overrun the first gap, so one must be shortened by a word; because the objective weights salience by narration time, the solver shortens the least salient of the three, selecting the first compressed variant of e_1 listed above ($y_{1,1} = 1$), while all other selected eventualities keep their full variant. The background details about the windows, road, mannequins and the tailor's appearance are dropped, as is the moderately salient e_{12}, for which no time remains.

[4] See `rottentomatoes.com/m/the_girl_with_the_dragon_tattoo/videos/ofdvX951xl2S`

Table 4: A prose scene description for a scene in *The Girl with the Dragon Tattoo* (2011). Occurrence is when each description is depicted on screen and salience is a score representing the relevance of each description to the story being told. Occ. is occurrence, and Sal. is salience. Scores are illustrative. d_i is the optimal delivery start time for selected eventualities; unselected eventualities are marked –.

	Occ.	**Sal.**	**Description element**	d_i
e_1	0s–4s	0.8	*Lisbeth stands with her hands on a table in an upmarket Stockholm tailor shop.*	0.0s
e_2	0s–4s	0.2	*The shop has large windows*	–
e_3	0s–4s	0.1	*that look out onto a road with parked cars,*	–
e_4	0s–4s	0.1	*and there are two mannequin torsos in front of the window.*	–
e_5	0s–4s	0.3	*A tailor is standing behind a large table.*	–
e_6	0s–4s	0.4	*He is a bald white man,*	–
e_7	0s–4s	0.2	*wearing a blue shirt and a brown vest.*	–
e_8	0s–2s	0.4	*The tailor drapes a garment bag across the table*	–
e_9	2s–4s	0.3	*and unzips it.*	–
e_{10}	4s–6s	0.85	*It contains a black, leather motorcycle jacket.*	4.0s
e_{11}	8s–10s	0.9	*It is identical to Mikael's jacket in a slide that Lisbeth is holding.*	6.1s
e_{12}	8s–10s	0.5	*The slide shows Mikael and Erika embracing.*	–
e_{13}	15s–17s	0.5	*The tailor zips the garment bag back up.*	–
e_{14}	17s–18s	0.7	*Lisbeth looks surprised by the tailor's comment,*	14.1s
e_{15}	17s–18s	0.6	*and gives him a wistful look.*	16.2s

The final decision variables are such that they maximize the objective function and meet the specified constraints. We can read these final decision variables off to generate the resulting AD script. For each eventuality i, if x_i indicates selection, then we read off the delivery time d_i and find the single y_{ik} that indicates selection. The textual description corresponding to y_{ik}, alongside its delivery start time represent the generated AD script.

B Evaluation Metrics

We summarize the REFRAMED metrics; see Sterner et al. (2026a) for full details. For the dialogue-gap metrics, gaps are intervals of at least one second between professional dialogue subtitles, and each reference and generated description is assigned to the gap that fully contains it, allowing a one-second collar. SODA-M and SODA-T are defined as

$$\text{SODA-M} = \frac{\sum_{(i,j)\in\mathcal{A}^\star} \text{METEOR}(r_i, \hat{r}_j)}{\frac{1}{2}\left(|\mathcal{R}| + |\hat{\mathcal{R}}|\right)}, \quad \text{SODA-T} = \frac{1}{|\mathcal{R}|} \sum_{(i,j)\in\mathcal{A}^\star} \mathbb{1}\Big[|\text{mid}(I_i) - \text{mid}(\hat{I}_j)| < \tau\Big]$$

where $\mathcal{R} = (r_i)$ and $\hat{\mathcal{R}} = (\hat{r}_j)$ are the reference and generated description sequences, I_i and $\hat{I}_j$ their time intervals, $\mathcal{A}^\star = \arg\max_{\mathcal{A}} \sum_{(i,j)\in\mathcal{A}} \text{METEOR}(r_i, \hat{r}_j)$ is their optimal monotonic alignment, $\text{mid}(\cdot)$ denotes time interval midpoint, and $\tau = 10$. QEval and QEval-T are defined as

$$\text{QEval} = \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}(\hat{a}_j = a_j)\,; \text{ QEval-T} = \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}(\hat{a}_j = a_j)\cdot T_j\,;\; T_j = \mathbb{1}(|\text{mid}(\hat{I}_j) - \text{mid}(I_j)| < \tau)$$

where N is the number of questions, a_j and $\hat{a}_j$ are the gold and predicted answers, and I_j and $\hat{I}_j$ are the reference and predicted attribution intervals.

Following Sterner et al. (2026a), metric scores are macro-averaged over sequences[5] within each movie, then over movies. In the permutation-based significance testing, we use sequences as the permutation unit.

[5] A sequence is a contiguous group of narratively related scenes. On the challenge set, sequences contain 6.8 scenes on average and there are on average 17.8 sequences per movie.

C SCENE DESCRIPTIONS

This appendix covers the scene descriptions from which the MILP's candidate eventualities are drawn: the prompt used to generate them from video (Appendix C.1), the character information supplied alongside the video so that characters are named correctly (Appendix C.2), and the screenplay scene descriptions used as an alternative source in our ablations (Appendix C.3).

C.1 LLM-GENERATED DESCRIPTION PROMPT

Scene Description Prompt

```
Your task is to generate a description of the visuals in the provided video (no
audio provided). The description should be written in the style of a scene
description from a screenplay, with your output as continuous narrative prose.

Adhere to all of the below.

Dos

- Write using complete sentences in the present tense and third person.
- Include only details explicitly depicted in the visuals.
- Follow the same order as the visuals.
- Write directly about the narrative world.
- Maintain objectivity rather than providing interpretations.
- Identify characters by name only if established by provided information,
otherwise by describing their physical appearance.
- Vary the density of the descriptions based on the amount of visual change.

Don'ts

- Do not output anything other than the prose description.
- Do not use meta-phrasing like video or scene.
- Do not mention cinematography details like shots or camera angles.
- Do not include audio details or conversational descriptors like speaking or
listening.
- Do not include repetitive details after they have been described once, such as
during a conversation without new visual information.

Once you have drafted your prose description, check that it adheres to all of the
above Dos and Don'ts. If it does not, revise it and check again. If it does,
respond with your description.
```

C.2 CHARACTER NAMING

Characters play a central role in movie narratives and it is essential that they are named correctly in AD. We apply two methods that support correct character naming: (1) subtitles are augmented by including the name of the character that speaks each line, (2) following Sterner et al.'s (2026a) task definition, LLMs that generate descriptions from video extracts alone are also provided with the names and faces of characters mentioned in the reference AD (overcoming the otherwise ambiguity in character naming). These methods are orthogonal in approach (one is video-based, the other text-based) and we expect that models can benefit from access to both.

Following Han et al. (2023b), we apply a face recognition tool for matching detected faces against (actor headshot, character name) pairs from IMDb. We also filter the recognized characters according to NER-identified characters for each video extract's AD (exact-match NER: $P = 97.2$, $R = 96.6$, $F_1 = 96.9$). Against the human-labeled test set from REFRAMED, our face recognition pipeline achieves $F_1 = 73.7$. Appendix C.2.3 gives more details. We treat the full-movie challenge set as a blind set, skipping the NER-based filtering (i.e., the reference AD transcripts are not used for any purpose except evaluation).

In terms of dialogue speaker identification, a scalable silver-standard is provided by the SDH in REFRAMED, because on occasions when the character speaking is not apparent from the visuals, the character's name is explicitly included (e.g. "[BOND]: The name is Bond, James Bond.") A second useful source of information is the Speechmatics speaker diarization provided by REFRAMED, which labels each detected dialogue with a speaker tag. We merge these two sources of information by propagating detected speaker names to all dialogue lines with the same speaker tag (Appendix C.2.1). Against a gold-standard provided by the 10 'post-production' screenplays, Speechmatics speaker diarization is highly effective at separating character speakers (confusion 12.8%, coverage 90.1%, purity 91.6%, outperforming all baselines, open-weight competitors and proprietary competitors we ran, Appendix Table 5). Combined with SDH-based speaker naming, we achieve $F_1 = 77.9$ on the speaker identification task (Appendix C.2.1).

C.2.1 Dialogue Speaker Identification

The task is to label each dialogue segment with the name of the character that is speaking. Our treatment of the task follows prior work (Korbar et al., 2024): first we rely on automatic speaker diarization that clusters dialogues with the same speaker, then we identify the character speaking a subset of dialogues, and finally we merge these two sources of information by identifying the character name associated with all dialogues in each cluster. To operationalize this, we rely on two newly available sources of information that result in a high-quality and scalable silver-standard: Speechmatics-based speaker diarization and SDH-based character naming.

Diarization. This step clusters dialogue by speaker identity. We investigate whether state-of-the-art tools can perform this task over the hours of dialogue in a movie. Our evaluation includes the open-weight models Pyannote 3.1 and Pyannote Community 1 as well as the proprietary systems Pyannote Precision 2 and Speechmatics Enhanced. Our baselines include a single speaker baseline that groups all speech into one cluster, a changing speaker baseline that assigns a unique cluster to every distinct speech segment, and a random oracle baseline that assigns each segment a random tag drawn from a fixed pool representing the true number of speakers.

Evaluation data is derived from the ten movies with best matched screenplays (see Appendix C.3). Performance is measured using confusion and missed detection rates, which reflect the proportion of dialogue duration assigned to an incorrect speaker and the proportion of undetected dialogue respectively. False alarm rates are omitted because our evaluation restricts analysis strictly to known gold dialogue segments. We additionally report cluster purity and coverage, which reflect the proportion of each predicted cluster belonging to a single gold speaker and the proportion of each gold speaker captured by a single predicted cluster respectively.

Table 5 gives results. All open-weight and proprietary systems outperform the baselines. The best open-weight system, Pyannote Community 1, achieves a confusion rate of 27.6 and a purity of 77.9. The proprietary systems perform better, and Speechmatics Enhanced achieves the lowest confusion at 15.8 and the highest purity at 89.7, alongside a coverage of 87.1 and a minimal missed rate of 0.1. REFRAMED provides manual annotation of the Speechmatics Enhanced cluster that corresponds to the audio description narrator. Removing all audio description dialogue reduces confusion further to 12.8 and increases purity to 91.6, while incurring only a marginal increase in the missed speech rate to 1.2.

Subtitle splitting Since our objectives are to apply the diarization to the professional subtitle provided by REFRAMED, we need to ensure that each subtitle entry corresponds to a single dialogue speaker. Professional movie subtitles typically adhere to guidelines in this respect: if more than one speaker is being included in a single subtitle entry, each speaker dialogue line is assigned to a separate line and each line is prefixed with a hyphen.

We identify subtitle entries containing multiple lines each prefixed with a hyphen. We attempt to split such entries into their individual lines by aligning their textual boundaries with word-level timestamps from the REFRAMED ASR transcripts. The split time is determined by finding an exact token match for either the final word of the preceding line or the initial word of the succeeding line within the transcript (after text normalization). Once these boundary times are identified, we assign new start and end timestamps to each extracted dialogue line. A minimum temporal gap of 0.024

Table 5: Character speaker diarization performance. Confusion is the proportion of dialogue duration assigned to an incorrect speaker; Missed Rate is the proportion of undetected dialogue. Both are computed after optimal one-to-one mapping between gold and predicted speaker tags. Purity is the proportion of each predicted cluster that belongs to a single gold speaker; Coverage is the proportion of each gold speaker captured by a single predicted cluster. Baselines: Single speaker assigns all dialogue to one tag; Changing speaker assigns a new speaker to every segment; Random Oracle K assigns each segment a random tag drawn from a fixed pool of the true number of speakers. The Final row removes all segments tagged as the AD narrator, where the narrator tag(s) are manually identified and the corresponding segments then automatically discarded (note the slightly higher Missed Rate).

System	**Confusion (↓)**	**Missed Rate (↓)**	**Purity (↑)**	**Coverage (↑)**
Baselines				
Single speaker	68.1	0.0	31.9	100.0
Changing speaker	95.6	0.0	100.0	4.4
Oracle K	89.0	0.0	33.8	11.7
Open-weight				
Pyannote Speaker Diarization 3.1	31.7	0.6	78.0	71.0
Pyannote Community 1	27.6	0.6	77.9	75.3
Proprietary				
Pyannote Precision 2	17.6	0.2	88.5	83.6
Speechmatics Enhanced	15.8	0.1	89.7	87.1
Final (AD semi-automatically removed)				
Speechmatics Enhanced	12.8	1.2	91.6	90.1

seconds is enforced between the resulting segments. We apply this to both the SDH and dialogue-only subtitles in REFRAMED.

Identification. Our speaker-identification pipeline operates in three stages. In the first stage, we identify the speaker of the segments that are apparent from the SDH. Our best methods rely on an LLM annotator (Olmo 3 32B Think) in order to extract the character names. For each SDH segment, we pass it as input to the LLM alongside contextual segments spanning 30 seconds on either side of the segment's start time (using SRT formatting). The model is instructed to return a character name only when the speaker is unambiguous (see page 19 for the prompt used). Subtitle segments formatted with hyphen prefixes on each line (multi-speaker exchanges) are split, when there exists a direct match for the first or last word of lines to be split with the REFRAMED ASR transcripts to use as the splitting time. Entries consisting only of bracketed non-speech content or remaining hyphenated lines are excluded before inference.

The second stage resolves each free-form name against the official cast list. A cascade matching procedure considers full names, all token subsequences, and set of common nicknames, mapping a name to a cast entry only when the form is claimed by exactly one cast member. Descriptive entries containing generic tokens such as "Man" or "Woman" are matched only by their full form, and age-prefixed entries are retained only when the base name appears separately in the cast.

The third stage transfers these labels onto speaker clusters from the ASR and speaker diarization transcripts. For each subtitle-level label we identify the cluster whose intervals cover the largest fraction of its duration and record a vote for that pairing. The cast entry receiving the most votes for a given cluster is taken as its label, after consulting an externally provided partial mapping from cast entries to canonical names and filling in any missing entries from the most frequent co-occurring free-form name. Each segment then inherits the canonical label of its cluster.

Evaluation. In terms of evaluation, the name used in the best-matching screenplay may not match the full name in the IMDb cast list, and the SDH-based names may not either. For evaluation, we resolve all names on both sides to the IMDb cast list entry using the cascade-based procedure. A

Table 6: Character speaker identification performance. The task is to identify the IMDb name of the speaker of each dialogue segment. Most frequent character assigns the modal reference character name to every segment; random IMDb character samples uniformly from the full IMDb cast; random reference character samples uniformly from character names in the reference; random-sampled reference character samples them in proportion to their reference frequency. Extracted segments only restricts predictions to segments the system named, leaving the rest unlabelled; speaker-based propagation assigns each speaker in Speechmatics diarization the system-predicted name that wins a majority vote across its segments, then labels every gold segment with the name attached to the diarised speaker it most overlaps with.

	P	R	F_1
Baselines			
Most frequent character	32.0	32.0	32.0
Random IMDb character	2.9	2.9	2.9
Random reference character	4.9	4.9	4.9
Random-sampled reference character	18.1	18.1	18.1
SDH extraction			
SDH, extracted segments only	89.8	3.3	6.4
+ speaker-based propagation	78.6	69.3	73.7
LLM prompting			
SDH, extracted segments only	82.2	4.1	7.9
+ speaker-based propagation	81.6	74.5	77.9

prediction is a true positive when its resolved cast entry matches the gold entry, a false positive when the entries differ, and a false negative when a gold segment is left unlabelled or mislabelled. We report micro-averaged precision, recall, and F_1.

We compare four baselines against two predictive systems, each in two configurations. The baselines (1) label every segment with the most frequent gold cast entry in the film, (2) sample uniformly from the full official cast, (3) sample uniformly from the set of cast entries in the gold labels, or (4) sample in proportion to the frequency of those entries. The predictive systems differ in how subtitle-level names are obtained. The first employs heuristics to identify common formatting SDH character-name cues (all-caps or title-case names preceding a colon and bracketed names at the start of a line). The alternative system uses LLM-based identification from SDH.

Results are in Table 6. The sampling-based baselines perform poorly ($F_1 = 2.9$, $F_1 = 4.9$, $F_1 = 18.1$). The strongest baseline labels every segment with the most frequent gold cast entry ($F_1 = 32.0$), unsurprising given that most movies have one or more main protagonists. The SDH-based approach outperforms all baselines on precision (extracting-only achieves $P = 89.8$ and 82.2 vs. the best baseline 32.0). Note that precision is imperfect here due to two factors: the SDH-based labels need to be propagated onto the dialogue-only subtitles, and resolving character names to the IMDb cast list entries can be noisy. Recall is low ($R = 3.3$ and $R = 4.1$; $F_1 = 6.4$ and $F_1 = 7.9$), demonstrating that only a small fraction of SDH segments explicitly cue the speaker. The LLM identifies more characters ($R = 3.3 \rightarrow 4.1$, $+0.8$) at a cost in precision ($P = 89.8 \rightarrow 82.2$, -7.6).

Speaker-based propagation improves recall dramatically ($R = 3.3 \rightarrow 69.3$ and $R = 4.1 \rightarrow 74.5$). LLM-based extraction achieves the best overall results, exceeding the extractive variant by a modest margin ($F_1 = 73.7 \rightarrow 77.9$, $+4.2$).

SDH Extraction Prompt

```
You are provided an excerpt of English movie subtitles spanning roughly one minute
around a target subtitle entry. Each subtitle entry has an index, a start timecode
and an end timecode.
```

```
My goal is to produce high-precision speaker-name labels for a subset of the
subtitle entries for a movie. A later stage will take as input all subtitle entries
and the high-precision labels for a subset of them, and handle the remaining
subtitle entries. Your task is to output the character name that speaks the target
subtitle entry when the speaker is unambiguous. When in doubt, output "unk".

Evidence can come from the dialogue or from non-dialogue information in the
subtitles. Subtitles sometimes add explicit character naming as a part of
non-dialogue information if the character is not visible on-screen. Whether each
character is visible is an independent factor. One character being named does not
imply the next character speaking will be named.

Rules:

1. Only commit to a character name when there exists direct evidence of the name
of the speaker of the target subtitle entry.
2. Evidence that applies to another subtitle entry cannot be used as evidence for
the target subtitle entry.
3. Do not attempt to guess which subtitle entries are spoken by the same character.
4. Do not infer the speaker from outside knowledge of the movie. Rely only on
evidence from the provided excerpt.
5. If your output is a character name, output the character name only, omitting
any other description, dialogue, etc.

INSERT_CONTEXT

Which character spoke the following target subtitle entry?

INSERT_TARGET

Your output should be a character name or "unk".
```

C.2.2 Character Names via Subtitle Identities and AD NER

AD requires both identification of the relevant on-screen characters and their consistent naming (e.g. "Commander Bond" as opposed to "James Bond"). IMDb provides movie credits, but characters are frequently listed according to their full name and this may not be how they should be referred to in AD. The earlier character names extracted from SDH can help bridge this gap. Still, SDH names might differ from the names used in our reference AD (note that here we only consider the character names as they are referred to in the US AD, ignoring the UK version).

Our first objective is thus to obtain a consistent mapping from IMDb character entries to the name variations used in the text, and ultimately to a single definitive name for our generation systems to use. Linking to IMDb also allows us to extract official actor headshots. SDH character names are extracted from the approach described in Section C.2.1. To extract names from the US AD reference, we apply Named Entity Recognition (NER)[6], extracting all `PER` spans. Against the REFRAMED human annotation, the model establishes excellent performance (exact-match $P = 97.2$, $R = 96.6$, $F_1 = 96.9$).

We attempt to resolve each SDH and AD-extracted name to exactly one IMDb character name. Our algorithm first attempts to match full names and all token subsequences. Unmatched names are then checked against a dictionary of English nicknames[7]. (Matching is after normalization according to Radford et al. (2023).) A mapping is only finalized if the name maps unambiguously to exactly one cast member.

To determine the final name for each character, we aggregate occurrence counts for each resolved name variation across the SDH and the AD, and use the name form with the highest combined frequency. Another output is a list of characters present in the AD for each video extract.

[6] `https://huggingface.co/dslim/bert-large-NER`

[7] `https://github.com/carltonnorthern/nicknames`

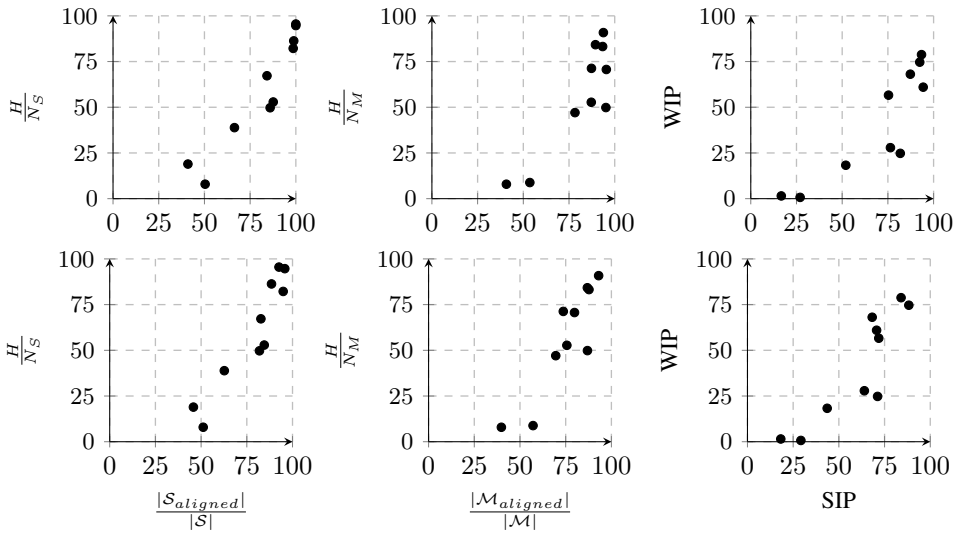


Figure 4: Screenplay–movie matching scores (WIP) against a measure of screenplay–movie scene alignment (SIP, defined below) for the 10-movie challenge set. The top row shows SIP metrics derived using human-annotated scene alignments, while the bottom row uses our automatic scene alignment method. The left and middle columns decompose the metrics into their directional components (representing deletion and insertion, respectively).

C.2.3 Character Face Detection

We address a simplified face recognition task that does not require full face tracking, but instead identifying one face for the character that is a good match. We apply automated face recognition to the video frames and, optionally, filter the resulting noisy predictions to those characters mentioned in the reference AD (according to the NER output). We sample video frames, detect all visible faces, and compare against the headshots using dlib-based face recognition.[8]. If a face receives more than one match, it is assigned to the actor to which it receives highest match.

For evaluation, we use the human-labeled test set from REFRAMED ($N = 230$ character faces). A prediction is a true positive if the predicted character name matches the gold label and the predicted face crop auto-matches the reference face. Our pipeline achieves high recall (71.7), but without NER-based filtering lower precision (46.7; $F_1 = 56.6$). After NER-based filtering, precision improves to 83.9, at the expense of a modest drop to recall (65.7). Overall the pipeline achieves a final $F_1 = 73.7$. Recall that since we treat the full-movie challenge set as a blind set, the NER-based filtering is skipped.

C.3 Screenplay descriptions

Quantifying Screenplay–Movie Match Our quantification of the degree of alignment between screenplay drafts and final film edits is based on matching the movie dialogue (from subtitles) with screenplay dialogue. Our metric is Word Information Preserved (WIP; Morris et al., 2004), a symmetric, bounded and information-theoretic metric that measures the mutual information between the two dialogue sources. We demonstrate the effectiveness of this metric by showing that its continuous scores correlate with the manual human annotation of movie–screenplay alignment from REFRAMED (Figure 4; Spearman $\rho = 0.87$, $p = 0.002$ with human movie–screenplay scene alignments, and $\rho = 0.83$, $p = 0.005$ with automatic alignment). More details follow the below description on QA evaluation.

Description quality-assessment with QA evaluation We evaluate the quality of LLM-generated scene descriptions using QEval on the validation set (i.e., answering AD-generated questions using

[8] https://github.com/ageitgey/face_recognition

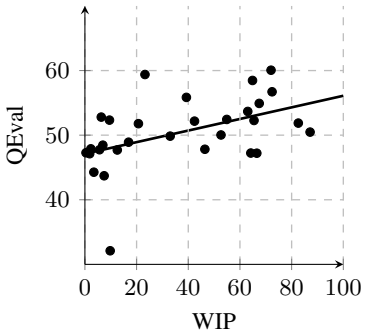


Figure 5: QEval scores using scene descriptions extracted from screenplays. Each point represents a single movie, and WIP is our measure of the degree of match between the movie's screenplay and the post-production movie.

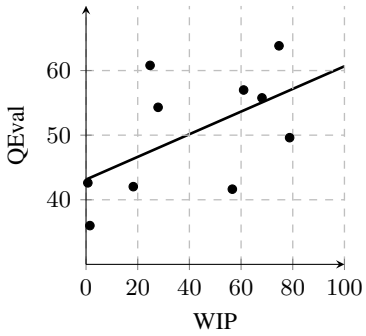


Figure 6: QEval scores against our measure of screenplay–movie match, WIP. The ten points represent each of the ten movies in the REFRAMED challenge set, which benefit from human-labeled screenplay alignment.

the generated scene descriptions). Reference scene descriptions extracted from the movie screenplay answer 52.9% of the questions correctly, demonstrating their relevance for the task. This result also shows that screenplays do not describe a substantial portion of the visual content included in audio description, a result of the flexibility afforded to movie directors and post-production editors. Using Gemini 3.1 Flash-Lite as an automatic LLM scene describer, we achieve 47.7% on the same task. Adding the automatic character faces improves performance to 48.6%, and further adding speaker timing results in 50.6%. These results demonstrate that LLMs approach the QA performance of screenplay descriptions. Results for 30 sampled movies in the REFRAMED dataset against WIP are shown in Figure 5; the relationship is moderate, Spearman $\rho = 0.52$, $p = 0.003$, with QEval rising by about 9 points across the full WIP range. Results for the ten manually aligned challenge-set movies are shown in Figure 6.

More on screenplay-movie scoring. Screenplays in the MovieSum dataset (Saxena & Keller, 2024), as used for REFRAMED, are from publicly available sources. They lack metadata of the date or draft version. The screenplays vary from early drafts to post-production versions. Since they can deviate from the final movie, it is of practical value to quantify the degree of alignment between the screenplay version in MovieSum and the final movie. We follow Lambert et al. (2013) by providing a continuous quantification, but unlike them stop short of classifying the resulting scores. Our continuous quantification is based on the information-theoretic formulation of the mutual information between two sequences operationalized by Morris et al. (2004), in particular word information preserved (WIP). Intuitively, between two sequences of words, WIP represents the probability that any given word from the first sequence is correctly matched with an identical word in the second sequence, and vice versa. This interpretable property makes it an attractive measure and is operationalisable by comparing the sequence of screenplay dialogue words against the sequence of words spoken in the final movie dialogue.

We compute an alignment that minimizes edit distance between screenplay dialogue and movie subtitle dialogue.[9] Let H be the number of matched words via minimum edit distance alignment, N_S be the total number of screenplay words, and N_M be the total number movie subtitle words. WIP is defined as:

$$WIP = \frac{H}{N_S} \cdot \frac{H}{N_M} \tag{2}$$

It multiplies match rates in both directions, and is hence a symmetric measure bounded in [0,1]. The first component $\frac{H}{N_S}$ represents the proportion of screenplay dialogue retained in the final film (penalizing deletions), while the second component $\frac{H}{N_M}$ represents the proportion of the final movie's dialogue that is also in the screenplay (penalizing insertions).

We validate the effectiveness of WIP as a continuous measure of screenplay-movie alignment by comparing against the human-annotated screenplay-movie alignment on our challenge set. The hu-

[9]Text is normalized according to Radford et al. (2023).

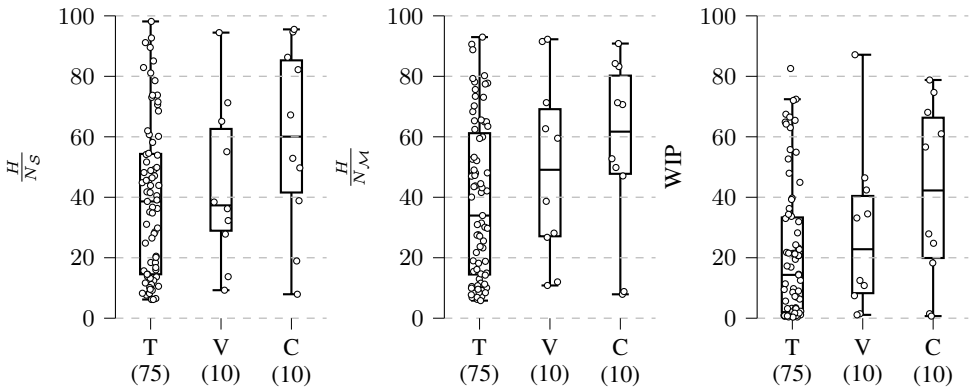


Figure 7: Distribution of WIP, and its two component parts (left two plots, representing screenplay deletion and insertion respectively), across dataset splits. T=Train, V=Validation, and C=Challenge splits; numbers in brackets represent the number of movies with aligned screenplays in each split.

man annotation is at the scene level, so we introduce scene information preserved (SIP) as a structural analogue to WIP. Let $|\mathcal{S}|$ and $|\mathcal{M}|$ be the total number of scenes in the screenplay and the movie. Our analogue for a word-level *hit* is for a screenplay scene to be aligned with at least one scene in the counterpart. Let $|\mathcal{S}_{aligned}|$ and $|\mathcal{M}_{aligned}|$ be the number of scenes with at least one aligned counterpart. SIP is formulated as:

$$SIP = \frac{|\mathcal{S}_{aligned}|}{|\mathcal{S}|} \cdot \frac{|\mathcal{M}_{aligned}|}{|\mathcal{M}|} \tag{3}$$

Again, these two components can be interpreted as deletion and insertion-based metrics, respectively, and (akin to WIP) SIP provides a metric bounded in [0,1] representing the degree of alignment. (Note that, similar to WER vs. WIP, the metric would not be bounded if we used raw edit distance.)

The top three plots in Figure 4 shows plots of SIP against WIP, as well as the insertion and deletion component parts against each other. We observe a positive correlation between WIP and SIP (top right plot, Spearman $\rho = 0.87$, $p = 0.002$). The bottom three plots represent results when we substitute ground-truth scene alignment with the REFRAMED automatic screenplay–movie scene alignment, where the same correlation holds (Spearman $\rho = 0.83$, $p = 0.005$).

We compute WIP scores for all screenplays included in the REFRAMED dataset. Figure 7 shows the resulting WIP for the training and validation splits (challenge data is also provided for reference; the REFRAMED test split is not provided with screenplays). The wide variance confirms the wide range of screenplay versions in the MovieSum dataset.

The movies with highest-WIP screenplay serve to provide a comparison source of scene descriptions for testing our AD generation methods (all from the training and validation sets). We designate the subset of eight movies with highest WIP as a REFRAMED post-production screenplay set.[10]

D LLM PROCESSING

The prompt below takes the video and the segmented scene description, and returns the MILP inputs for each description element: its occurrence span, its five compressed variants, and its salience score (Section 4).

[10]We selected the top ten, but after manual inspection, two of the movies were later excluded: *The Breakfast Club (1985)*, due to its screenplay being incorrectly parsed in MovieSum (scene descriptions are inside dialogue tags), and *Tár (2022)*, also due to the parsing in MovieSum (scene segmentation is not provided). Note that in both cases, the screenplays are both of high quality (in correspondence with WIP), with movie dialogue a near-perfect match against dialogue in the screenplay.

Candidate generation prompt

```
Task description:

You are provided with the video for a movie scene, alongside a prose scene
description. Your task is to provide temporal grounding for each extract of the
scene description, and different formulations of it that could be used as audio
description.

Audio description is a verbal description of key visual content in a movie. It
provides blind and visually impaired individuals access to information necessary
for following the story.

Output format:

Return exactly one valid JSON array and no other text. Each element represents a
single extract from the prose scene description. The following illustrates the
required structure for each element.

{"scene_description_extract": "SCENE_DESCRIPTION_EXTRACT", "audio_description":
"DESCRIPTION", "compressed_audio_descriptions": {"0.9": "DESCRIPTION", "0.8":
"DESCRIPTION", "0.7": "DESCRIPTION", "0.6": "DESCRIPTION", "0.5": "DESCRIPTION"},
"occurrence_start": "OCCURRENCE_START", "occurrence_end": "OCCURRENCE_END",
"salience": SALIENCE}

- SCENE_DESCRIPTION_EXTRACT is the textual extract from the prose scene
description.
- DESCRIPTION is a formulation of the same visual content as the extract that can
be used as audio description.
- OCCURRENCE_START and OCCURRENCE_END indicate the temporal span in which the
described visual content is established for a sighted viewer.
- SALIENCE represents the extent to which the visual content provides information
necessary for following the story, relative to the other candidates in the output
array. It must be a float greater than 0.0 and less than 1.0.

Rules for each element:

- SCENE_DESCRIPTION_EXTRACT must remain identical to the provided extract.
- The "audio_description" should contain a copy of the scene description extract,
reformulating only as needed to form a natural audio description.
- The "compressed_audio_descriptions" dictionary must contain exactly the keys
"0.9", "0.8", "0.7", "0.6", and "0.5", which refer to target compression rates of
the "audio_description".
- The value for each key should target approximately that proportion of the word
count of the "audio_description". Punctuation does not count as words.
- All generated descriptions must be a single sentence.
- OCCURRENCE_START and OCCURRENCE_END are timestamps with format
"TIMESTAMP_FORMAT".
- OCCURRENCE_START >= 0.
- OCCURRENCE_END > OCCURRENCE_START.
- If the visual content described by an extract does not occur in the video or
cannot be visually established, set OCCURRENCE_START, OCCURRENCE_END, and
SALIENCE to null.

Occurrence spans for different extracts may overlap.

The following is the required output template, with the required scene description
extracts included.

OUTPUT_TEMPLATE
```

```
Generate the JSON with the "audio_description", "compressed_audio_descriptions",
"occurrence_start", "occurrence_end", and "salience" values filled in. Your JSON
must include all elements in the template above.
```

E LLM SCHEDULING

The prompt below is used for the Qwen and Gemini scheduler baselines (Section 4), which replace the MILP solver with an LLM given the same candidate elements and dialogue gaps.

LLM Scheduling Prompt

```
Task description:

You are provided with information about a movie scene. The information is
represented in two JSON arrays: (1) candidate elements containing descriptions,
temporal grounding, salience scores, and narration durations; (2) the start and
end times of dialogue gaps. Your task is to select a subset of the elements and
assign them delivery times. The selected elements will be narrated at the assigned
delivery times as audio description for the movie scene.

Audio description is a verbal description of key visual content in a movie. It
provides blind and visually impaired individuals access to information necessary
for following the story.

Input format:

You will receive two JSON arrays.

The first array contains candidate elements with the following structure:

{
  "element_id": "UNIQUE_ELEMENT_ID",
  "description": "DESCRIPTION_TEXT",
  "occurrence_start": NUMBER,
  "occurrence_end": NUMBER,
  "salience": NUMBER,
  "duration": NUMBER
}

The second array contains dialogue gaps with the following structure:

{
  "dialogue_gap_start": NUMBER,
  "dialogue_gap_end": NUMBER
}

All times and durations are expressed in seconds.

Output format:

Return exactly one valid JSON array and no other text. Each entry selects one
element and assigns it a delivery time.

[
  {
    "element_id": "UNIQUE_ELEMENT_ID",
    "delivery_start": NUMBER
  }
]
```

```
Rules:

- The entries in your generated JSON array must contain only the keys "element_id"
and "delivery_start".
- Each "element_id" must be in the JSON array of candidate elements.
- Each "element_id" may be selected at most once.
- Express "delivery_start" times in seconds.
- You may select no elements by returning [].

JSON array of elements:

ELEMENTS_JSON_ARRAY

JSON array of dialogue gaps:

DIALOGUE_GAPS_JSON_ARRAY
```

F SYSTEM OUTPUT FOR THE SOCIAL NETWORK (2010)

The AD references shown in this section come from the REFRAMED training split, where AD is automatically transcribed rather than professionally transcribed as in the challenge set (Section 4). Visible errors, for instance "Christy *sense* the scarf alight" for *sets*, or "Christy *drugs* . . . Eduardo into a toilet cubicle" for *drags*, are artifacts of that transcription, not errors by the describer. We reproduce the references verbatim, exactly as they were scored, rather than correcting them. This is one reason our main results are reported on the professionally transcribed challenge set, and a further reason to treat n-gram overlap against these particular references with caution. In these examples, MILP refers to the fully automatic Qwen MILP system, and MILP (screenplay) refers to the variant that uses scene descriptions extracted from a screenplay.

F.1 PUTTING OUT FIRES

Why does your Status say "single" on your Facebook page?
[+3s] **What?**
[+4s] **Why does your Relationship Status say "single" on your Facebook page?**
[+8s] **Well, I was single when I set up the page.**
[+10s] **And you just never bothered to change it?**
[+12s] **What?**
[+13s] **I don't know how.**
[+15s] **Do I look stupid to you?**
[+17s] **No, calm down.**
[+18s] **You're asking me to believe that the CFO of Facebook**
[+21s] **doesn't know how to change his Relationship Status on Facebook?**
[+23s] **It's embarrassing, so you should take it as a sign of trust that I would tell you that.**
[+26s] **- Go to hell. - Take it easy.**
[+28s] **No, you didn't change it so you could screw those Silicon Valley sluts**
[+31s] **every time you go out to see Mark.**
[+32s] **Not even remotely true, and I can promise you that the Silicon Valley sluts**
[+35s] **don't care what anyone's Relationship Status is on Facebook.**
[+38s] **Please, open your present.**
[+40s] **Your phone does work.**
[+43s] **It's Mark.**
[+44s] **Okay, this is gonna be tricky.**
[+46s] **Open your present. It's a silk scarf.**
[+49s] **Have you ever seen me wear a scarf?**
[+51s] **This'll be your first.**
[+53s] **Yeah.**
[+54s] **- You froze our account? - I did.**
[+55s] **- You froze the account. - I had to get your attention, Mark.**
[+57s] **Do you realize that you jeopardized the entire company?**
[+59s] **Do you realize that your actions could have destroyed**
[+61s] **everything I've been working on?**
[+62s] **We have been working on.**
[+63s] **Without money, the site can't function.**
[+65s] **Let me tell you the difference between Facebook and everybody else.**
[+67s] **We don't crash ever!**
[+68s] **If the servers are down for even a day, our entire reputation is irreversibly destroyed.**
[+72s] **- Look... - Users are fickle. Friendster has proved that.**
[+74s] **Even a few people leaving would reverberate through the entire user base.**
[+77s] **The users are interconnected. That is the whole point.**
[+79s] **College kids are online because their friends are online**
[+81s] **and if one domino goes, the other dominos go.**
[+83s] **Don't you get that?**
[+85s] **I am not going back to the Caribbean Night at A-E-Pi!**

American AD: [+87s] *Christy lights her gift on fire.*
British AD: [+87s] *Christy sense the scarf alight.*

MILP: [+87s] *The woman holds a lighter*
MILP (screenplay): [+87s] *She lights the box.*

[+88s] **Holy shit.**

American AD: [+89s] *She drops it into a trash can,* [+91s] *then tips it onto his bed.*
British AD: [+91s] *and drops it in a bin.*

MILP: [+90s] *Flames erupt from the wastebasket on the bed*
MILP (screenplay): [+90s] *He then tosses the phone down onto the bed.*

[+92s] **What is wrong with you?**
[+94s] **Did you like being nobody?**
[+95s] **Did you like being a joke? Do you wanna go back to that?**
[+97s] **Hang on, hang on, hang on, hang on.**
[+99s] **That was the act of a child, not a businessman,**
[+100s] **and it certainly was not the act of a friend.**
[+102s] **You know how embarrassing it was for me to try to cash a check today?**
[+104s] **I am not going back to that life.**
[+106s] **- Maybe you were frustrated. - Yeah!**
[+108s] **- Maybe you were angry. - I was!**
[+111s] **But I am willing to let bygones be bygones,**
[+113s] **because, Wardo, I've got some good news.**
[+116s] **I'm sorry.**
[+117s] **I was angry, and maybe it was childish, but I had to get your attention.**
[+122s] **Wardo, I said I've got some good news.**
[+125s] **What is it?**
[+126s] **Peter Thiel just made an angel investment of half a million dollars.**
[+129s] **What?**
[+130s] **Half a million dollars. And he's setting us up in an office.**
[+135s] **They wanna reincorporate the company. They wanna meet you.**
[+137s] **They need your signature on some documents,**
[+139s] **so you gotta get your ass on the first flight back to San Francisco.**
[+141s] **I need my CFO.**

American AD: [+143s] *Eduardo squeezes his eyes shut,* [+144s] *swallows* [+145s] *and shakes his head.*
British AD: [+143s] *Having extinguished the fire, Eduardo closes his eyes* [+146s] *and shakes his head.*

MILP: [+143s] *He turns around to look behind him* [+145s] *The woman stands in the doorway watching him closely*
MILP (screenplay): [+144s] *There is a quiet pause for a moment.*

[+148s] **I'm on my way.**
[+149s] **- Wardo? - Yeah?**
[+151s] **We did it.**

American AD: [+152s] *He tearfully snaps his phone shut.*
British AD: [+152s] *Eduardo hangs up* [+153s] *and Christy appears in the doorway.* [+156s] *Where are.*

MILP (screenplay): [+152s] *Eduardo jumps in surprise as Christy stands directly behind him.*

[+156s] **Wardo?**
[+157s] **You're going back there already?**
[+159s] **Yes.**

MILP (screenplay): [+161s] *The scene cuts away.*

Sampled QA: What does Eduardo do immediately after extinguishing the fire?
A. calls someone B. starts laughing **C. closes his eyes** D. leaves the room E. opens the door

F.2 I'M NOT A BAD GUY

We're done for the day.
[+2s] **Yeah. Yeah, I was just sitting here.**

American AD: [+5s] *He faces an open laptop* [+6s] *and punches a few keys.*
British AD: [+5s] *Everyone else has gone.*

MILP: [+4s] *Marylin enters the conference room holding a light jacket.* [+7s] *Typing on laptop.*
MILP (screenplay): [+4s] *Mark types on a laptop at a table.* [+7s] *She holds a coat.*

[+8s] **What happened to Sean?**
[+10s] **He still owns 7% of the company.**

American AD: [+13s] *He turns to face the young attorney in* [+14s] *his melancholy scowl fades.*
British AD: [+13s] *She nods.* [+14s] *Mark looks up from his laptop.*

MILP: [+13s] *She walks to the room's middle.* [+15s] *Mark halts his typing activity abruptly.*
MILP (screenplay): [+13s] *Mark looks up from his computer at Marylin.*

[+17s] **All you had all day was that salad. Do you wanna get something to eat?**
[+21s] **I can't.**

American AD: [+22s] *Mark nods understandingly* [+24s] *and briefly works his mouth.* [+25s] *He turns away, shyly,* [+26s] *shuts his laptop,* [+27s] *then faces her again.*
British AD: [+24s] *Mark nods and looks thoughtful.* [+26s] *He opens his mouth to speak,* [+27s] *then sighs and closes his laptop.*

MILP: [+23s] *He lifts his eyes upward.* [+24s] *He makes an emphatic gesture using his right hand.* [+27s] *Then he fully shuts the laptop lid downward.* [+29s] *Rests hands now.*

[+30s] **I'm not a bad guy.**
[+32s] **I know that.**
[+33s] **When there's emotional testimony, I assume 85% of it is exaggeration.**
[+37s] **And the other 15?**
[+39s] **Perjury. Creation myths need a devil.**

American AD: [+42s] *Mark sighs* [+43s] *and drums his laptop with his fingers.*
British AD: [+42s] *Mark drums his fingers on his laptop.*

MILP: [+42s] *Marylin remains positioned at the table's head observing.* [+44s] *Shifts to side.*

[+45s] **What happens now?**

MILP: [+47s] *Occupies vacant seat.*

[+48s] **Sy and the others are having a steak on University Avenue.**
[+53s] **Then they'll come back up to the office,**
[+54s] **and start working on a settlement agreement to present to you.**
[+58s] **- They're gonna settle? - Oh, yeah.**
[+60s] **- And you're gonna have to pay a little extra. - Why?**
[+63s] **So that these guys sign a nondisclosure agreement.**
[+65s] **They say one unflattering word about you in public, you own their wife and kids.**
[+68s] **I invented Facebook.**
[+70s] **I'm talking about a jury.**
[+72s] **I specialize in voir dire, jury selection.**
[+75s] **What a jury sees when they look at a defendant.**
[+77s] **Clothes, hair, speaking style, likeability...**
[+80s] **Likeability.**
[+81s] **I've been licensed to practice law for all of 20 months,**
[+84s] **and I could get a jury to believe**
[+85s] **that you planted the story about Eduardo and the chicken.**
[+88s] **Watch what else.**
[+90s] **Why weren't you at Sean's sorority party that night?**
[+93s] **- You think I'm the one that called the police? - Doesn't matter.**
[+95s] **I asked the question, now everybody's thinking about it.**
[+97s] **You've lost your jury in the first 10 minutes.**

American AD: [+99s] *Mark taps his laptop.*

MILP: [+99s] *Positional shift made.*

[+101s] **Farm animals.**
[+102s] **Yeah.**

MILP: [+103s] *Bends ahead gently.*

[+104s] **I was drunk and angry and stupid.**
[+107s] **- And blogging. - And blogging.**
[+110s] **Pay them.**
[+112s] **In the scheme of things,**
[+113s] **it's a speeding ticket.**
[+117s] **That's what Sy will tell you tomorrow.**
[+119s] **Do you think anybody would mind if I stayed and used the computer for a minute?**
[+122s] **I can't imagine it would be a problem.**
[+124s] **Thanks.**
[+126s] **I appreciate your help today.**

American AD: [+128s] *As Mark opens his laptop, the young attorney pauses and turns back.*
British AD: [+128s] *Marylin strolls to the glass door behind him* [+131s] *and turns.*

MILP: [+128s] *After extended silence, Marylin rises from her chair.*
MILP (screenplay): [+128s] *Marylin picks up her briefcase.*

[+132s] **You're not an asshole, Mark.**

American AD: [+134s] *He faces her with a furrowed brow.*

MILP: [+134s] *Advances to doors.*

[+136s] **You're just trying so hard to be.**

American AD: [+137s] *Marylin regards him sympathetically,* [+139s] *then turns away abruptly* [+140s] *and pushes out through a glass door.* [+142s] *Shouldering her bag, she crosses a reception area outside.* [+146s] *Mark lingers at the conference table, staring distantly.* [+149s] *He shifts his gaze,* [+150s] *blinks,* [+150s] *then eyes his laptop screen.*
British AD: [+139s] *He gazes at her over his shoulder.* [+141s] *Marylin turns and leaves,* [+143s] *slipping her handbag over one shoulder as she trots out through the adjacent office.* [+148s] *Mark looks thoughtful.*

MILP: [+138s] *Marylin halts momentarily at the doorway entrance.* [+140s] *She turns her head to observe the interior.* [+143s] *Marylin departs the conference space entirely.* [+144s] *Guard observed.* [+145s] *Mark stays isolated at the conference table.* [+147s] *Mark lifts the laptop display to reactivate it.* [+150s] *He recommences data input on the keyboard.* [+152s] *A fleeting screen glance precedes a pensive expression.*
MILP (screenplay): [+138s] *She walks out of the room.* [+142s] *Mark settles in at his laptop.* [+146s] *Mark logs onto his computer.* [+148s] *Mark smiles at the screen.* [+150s] *and waits for the response.*

Sampled QA: What does Mark do after turning to face the young attorney?
A. waves B. smiles C. shrugs D. frowns **E. nods**

F.3 We Have Groupies

So what were their names?

MILP (screenplay): [+2s] *The scene transitions to somewhere.*

[+4s] **Their names were Christy and Alice,**

MILP: [+6s] *Mark sits at a wooden table, looking serious.*
MILP (screenplay): [+6s] *Suddenly, a wooden stall door swings open.*

[+8s] **and they wanna have drinks tonight.**

American AD: [+11s] *Now, in a public bathroom.*

MILP: [+11s] *While Eduardo leans in close, facing him.*
MILP (screenplay): [+11s] *Eduardo is forcefully shoved into the stall.*

[+13s] **You're not supposed to be in here. This is a men's room.**

American AD: [+16s] *She rips her shirt off* [+18s] *in Eduardo eyes.* [+19s] *Her body.*
British AD: [+17s] *Christy drugs.* [+18s] *Eduardo into a toilet cubicle at a fancy club.*

MILP: [+16s] *The setting changes to a low angle in a bathroom.*
MILP (screenplay): [+16s] *Christy follows him in, having pushed him inside.* [+19s] *She leans close to him.*

[+21s] **Wow.**

American AD: [+21s] *Kissing him again, she clutches his head and neck.* [+28s] *Down low on the tile floor, we glimpse Mark and Alice feet in the next stall.*
British AD: [+22s] *She kisses him with hungry abandon.* [+24s] *Hearing a noise, he pulls away.* [+36s] *Mark.*

MILP: [+22s] *Where a woman's legs in black boots walk past a stall partition.* [+26s] *Above, a young man and the woman kiss intensely.* [+28s] *She holds his head.* [+29s] *While he leans back against the wall.* [+32s] *The view shifts underneath the stall.*
MILP (screenplay): [+22s] *She pins him against the wooden stall divider.* [+24s] *Eduardo's hands slide underneath Christy's white shirt.* [+27s] *His hands find her red bra just as they hear a noise.* [+30s] *Someone has just entered the adjacent stall.* [+32s] *Kept pinned tight.* [+34s] *She reaches down to unbuckle his belt.*

[+36s] **Oh, my God.**
[+38s] **I don't care.**

American AD: [+39s] *They kiss again* [+40s] *and she unbuckles his belt,* [+42s] *unzips his pants* [+43s] *and reaches in.* [+44s] *The view down low to the floor shows Marc's pants drop to now.* [+48s] *Christy kisses her way down Eduardo's chest.* [+50s] *We slowly ascend the dark wood of the stall doors,* [+54s] *later standing outside the bathroom.* [+57s] *A happy Eduardo glances at his friend.* [+59s] *A guy approaches.*
British AD: [+40s] *Christy feverishly unbuckles his belt* [+42s] *and thrusts her hand down his boxer shorts* [+45s] *in the adjacent cubicle.* [+46s] *Alice tugs Mark's jeans down.* [+49s] *Christy kisses Eduardo's chest* [+51s] *as she sinks to her knees.* [+53s] *Eduardo and Mark Stand guard outside the restroom door.* [+56s] *Eduardo grins inanely at Mark, who stares into space.* [+59s] *A guy approaches.*

MILP: [+40s] *Showing the woman standing beside the seated man among scattered shoes and bags.* [+43s] *A close-up reveals hands adjusting a belt buckle before the man stands up, straightening his coat.* [+48s] *Later, Mark and Eduardo stand together in a room lit by candles.* [+52s] *Smiles exchanged.* [+55s] *They turn their attention to a glass window.*
MILP (screenplay): [+40s] *Another noise comes from the stall next to them.* [+42s] *Christy has now unzipped Eduardo's fly.* [+44s] *Eduardo looks down through the gap between the stalls.* [+47s] *He sees a pair of Adidas sneakers on the floor.* [+50s] *Before he speaks.* [+51s] *Christy pulls her shirt open to reveal the red bra.* [+54s] *She places her hand inside his pants before the scene cuts.* [+57s] *Mark and Eduardo stand silently outside the bathroom door.* [+60s] *They exchange quiet, happy glances.*

[+62s] **Hey, man, sorry.**
[+64s] **A couple girls are freshening up in there.**
[+67s] **Sweet.**

American AD: [+68s] *They share a grin* [+69s] *and he goes off.* [+70s] *Eduardo smile lingers.*
British AD: [+68s] *Eduardo nods bashfully* [+70s] *and grins from ear to ear.* [+71s] *The guy goes.*

MILP: [+68s] *Where a man in a beanie grins at them.*
MILP (screenplay): [+68s] *A man approaches the door intending to use the restroom.* [+71s] *Gives up leaves.*

[+73s] **We have groupies.**

American AD: [+74s] *They swap a grin.* [+78s] *Peering across the busy restaurant, Mark sees Erica with some friends.*
British AD: [+74s] *Mark cracks a smile,* [+75s] *and they share a little.* [+79s] *Mark sees somebody in a smile fade.* [+80s] *He leaves Eduardo and guard duty.*

MILP: [+74s] *Mark and Eduardo peer into a busy restaurant filled with diners at candlelit tables.*
MILP (screenplay): [+74s] *Taps Mark on arm.* [+75s] *Mark finds himself smiling despite the awkward situation.* [+78s] *Mark's expression changes as he spots something.* [+80s] *Mark navigates through the crowd towards a booth.*

[+83s] **I'll be right back.**
[+84s] **Mark, where you going? Mark.**

American AD: [+86s] *He watches uneasily as Mark approaches his ex-girlfriend's table.* [+91s] *Erica.*
British AD: [+86s] *Mark crosses the candlelit room* [+88s] *and approaches a table where Erica sits with friends.*

MILP: [+86s] *A girl wearing a dark beret sits at a table with a group.* [+90s] *Smiles while looking up.*
MILP (screenplay): [+87s] *A girl is seated at the booth.* [+89s] *Although her back is turned, Mark recognizes her.*

[+91s] **Erica?**

American AD: [+91s] *Erica.* [+92s] *She looks up,* [+93s] *then eyes him flatly.*
British AD: [+93s] *She eyes him.*

MILP: [+92s] *Stops awkwardly near table shifting weight.*
MILP (screenplay): [+93s] *Surrounded by men woman.*

[+94s] **Hi.**
[+95s] **I saw you from over there. I didn't know you came to this club a lot.**
[+98s] **- First time. - Mine, too.**
[+100s] **Could I talk to you alone for a second?**
[+103s] **I think I'm good right here.**
[+104s] **I just... I'd love to talk to you alone if we could just go someplace.**
[+108s] **Right here is fine.**
[+110s] **I don't know if you heard about this new website I launched.**
[+112s] **No.**
[+114s] **- The Facebook? - You called me a bitch on the Internet, Mark.**
[+117s] **That's why I wanted to talk to you.**
[+119s] **- On the Internet. - That's why I came over.**
[+121s] **Comparing women to farm animals.**
[+123s] **I didn't end up doing that.**
[+125s] **It didn't stop you from writing it.**
[+127s] **As if every thought that tumbles through your head was so clever**
[+129s] **it would be a crime for it not to be shared.**

American AD: [+131s] *Eduardo looks on.*
British AD: [+132s] *The groupies emerge.*

MILP: [+132s] *Mark glances back at Eduardo.*
MILP (screenplay): [+132s] *But it hurts Mark deeply.*

[+133s] **The Internet's not written in pencil, Mark, it's written in ink,**
[+136s] **and you published that Erica Albright was a bitch,**
[+139s] **right before you made some ignorant crack about my family's name, my bra size,**
[+143s] **and then rated women based on their hotness.**

MILP: [+145s] *Gaze returns to girl.*

[+146s] **- Erica, is there a problem? - No, there's no problem.**

MILP: [+149s] *More emphatic gestures.*

[+150s] **You write your snide bullshit from a dark room**
[+152s] **because that's what the angry do nowadays.**
[+155s] **I was nice to you. Don't torture me for it.**
[+158s] **If we could just go somewhere for a minute...**
[+160s] **I don't wanna be rude to my friends.**
[+162s] **Okay.**

American AD: [+164s] *He goes off.*
British AD: [+165s] *He turns away.*

MILP: [+164s] *Mark eventually turns.*
MILP (screenplay): [+164s] *Turns away walks off.*

[+166s] **Good luck with your video game.**
[+169s] **Hey, that was great. That was the right thing to do.**

[+171s] **You apologized, right?**

American AD: [+172s] *Mark glances aside.*
British AD: [+172s] *Mark's eyes rove.*

MILP: [+173s] *And walks away down a corridor.* [+174s] *By Eduardo.*
MILP (screenplay): [+173s] *He passes Eduardo and Christy, who are watching him.*

[+175s] **We have to expand.**
[+177s] **Sure. Mark?**

American AD: [+178s] *As Mark walks out, Eduardo faces the girls.* [+181s] *Something*
British AD: [+177s] *Mark* [+178s] *Mark storms out of the club.*

MILP: [+178s] *Two women observe scene from hall end.*
MILP (screenplay): [+178s] *Mark exits the establishment through the front door.*

[+180s] **Is he mad about something?**

Sampled QA: Who is sitting with Erica at the table?
A. Mark B. stranger C. Eduardo D. guy **E. friends**